nature

www.nature.com/nature

Vol 445 | Issue no. 7123 | 4 January 2007

Avoiding a chimaera quagmire

Researchers need to take the initiative in addressing a controversial and urgent ethical issue: under what circumstances should the fusion of cells of animals and humans be permitted?

Last month, legislative revisions to the UK Human Fertilisation and Embryology Act were proposed in which the British government sought to reflect developments in reproductive technology and research since the act's inception in 1990. Among the recommendations (see www.dh.gov.uk/assetRoot/04/14/13/15/04141315.pdf) are new regulations on embryo research combining human and non-human material that take into account acute public concerns over where such research might lead.

This is set to ignite a debate about a central approach in the transition from basic research on stem cells to their clinical use, which might eventually reduce the need for human eggs and embryos in research. The onus is now on the scientific community to assume a proactive role in helping to define the boundaries of what is ethical and safe in interspecies research using human cells. At stake are both public support and the risk of delays to high-priority biomedical research.

In Britain, the governance of research combining cells, cellular material or genetic material from different individuals or species is nebulous. The new proposals provide a well-reasoned and permissive approach to the contentious issue of the nuclear transfer of human nuclei to a human donor egg for the purposes of therapeutic cloning, and also propose that other research combining animal and human materials should be subject to more stringent licensing procedures. The devil will be in the details, or lack of them, and will have policy implications for scientists working on cell-replacement therapies.

The document recognizes that human–animal fusion products have been widely used in biomedical research for many years (for example, in xenograft models of cancer in which human cells are introduced into mice), but also notes public unease with them. The lack of specific language in the document highlights a gap in the regulations that will need to be filled. How researchers will influence such policy decisions depends on their ability to swiftly agree and promote key research priorities, coupled to a clear assessment of how these priorities are balanced by both safety and ethical concerns.

Current restrictions

In the United States, the National Institutes of Health does not fund research involving the transplantation of human embryonic stem cells into animal embryos. In Canada, this funding restriction extends to the transplantation of human tissue-specific (or 'adult') stem cells to animal embryos. In 2005, the US National Academy of Sciences stated its opposition to research in which human embryonic stem cells are introduced into non-human primate blastocysts (pre-implantation embryos), or in which any embryonic stem cells are introduced into human blastocysts, as well as the breeding of any animal into which human embryonic stem cells have been introduced.

At present, such guidelines are reasonable but do not consider several promising and arguably necessary avenues of research that combine human cells or cellular components with other species. These include combining the genetic material of humans and other species, the prenatal combination of cells from different individuals (animal to human, human to animal, or human to human), or grafting tissue from humans to animals.

One troubling outcome of a debate could be UK restrictions on current work combining factors from animal eggs (which, unlike human eggs, are readily accessible) with animal or human nuclei. These 'nuclear reprogramming' experiments aim to identify components of the egg that are capable of transforming an adult cell into one with the vast capabilities of stem cells. They could generate stem cells and tissues genetically and immunologically matched to patients, and obviate the need for human eggs and embryos in generating human embryonic stem cells. There are strong arguments for permitting such research, given the minimal safety risk or violation of human dignity when any resulting embryos are arrested at an early stage.

A question of balance

Another important line of investigation is the engraftment of human stem cells into non-human primate models. Such work is essential, as it would be dangerous to have clinical trials for cell-replacement therapies in humans without first testing promising human cell lines in animal models, for example by transferring human neuronal cells into animal brains. A valuable and provocative discussion by Karpowicz *et al.* (*Nature Med.* **10,** 331–335; 2004) outlines some key questions in balancing ethics with human benefits: how robustly the transplanted cells are incorporated into the host; at what stage and into what tissues and organs they are introduced; whether there is a possibility that introducing such cells would alter the production of sperm or eggs in the host animal; and, for neuronal transplantation, whether there is a risk of transferring human functions or behaviours to the host animal. This would be an unacceptable outcome whose risk, Karpowicz *et al.* argue, is generally negligible.

The avenues of research discussed here all fall under the rubric of combining animal and human material, and offer a high level of benefit, but with disparate degrees of risk, so they should be regulated differently. Scientists must provide recommendations on how to proceed, on where there is consensus, and where significant risks or ethical problems exist.

Scientists should identify the various research protocols defining interspecies research involving human cells and embryos, and the associated risks, ethical issues and benefits of each. They should put forward clear and comprehensive recommendations to the public and to regulatory bodies. If they don't, they risk having regulation and funding restrictions imposed on their research that are out of proportion to the ethical or safety risks involved. Even worse, they could face prohibitions that lump together research with vast disparities in intent and in the balance of risk and benefit — ultimately penalizing those who stand to gain from the therapies that might emerge. ■

Libya and human values

Death sentences issued by a Libyan court highlight more than one type of injustice.

The huge international outcry that followed last month's unjust decision by a Libyan judge to sentence six health professionals to death is hardly surprising. The charge that they deliberately infected more than 400 children with HIV in 1998 was baseless. The authorities ignored a body of evidence demonstrating that the cause of the outbreak was the use of contaminated medical material in the hospital in Benghazi, and that many of the children were infected long before the medics even began working at the hospital (see page 7).

Libya has responded vigorously to the international community's reassertion that it should ensure a fair and impartial trial, and for scientific evidence to be taken in account. Its foreign ministry has denounced Western political interference as creating a dangerous precedent in which Libyans are considered "sub-human" and treated differently from Bulgarians. It added that the political stance expressed by the Bulgarian government, European Union countries and others shows "a clear bias to certain values that are likely to trigger wars, conflicts and cause enmity between religions and civilizations".

It would be too simplistic to dismiss this entirely as anti-Western rhetoric. There is understandable resentment in many parts of the world that powerful nations are selective and inconsistent in their application of human rights. But the attention attracted by the Libyan scandal has been largely fuelled by the social conscience of what can in such instances be justifiably called the international scientific community — a force that is largely apolitical. It has a long track record in defending individuals on trial in human-rights cases, and has helped Arab and other scientists who have suffered travel restrictions to the United States (see *Nature* **443,** 605–606; 2006). It has also been relatively even-handed in its struggle to champion science as a rational means of establishing truth, and has consistently attacked the abuse of science for political ends wherever this occurs.

The case of the health professionals is an eminently scientific one, and the protests of the global scientific community are a defence not of Western values, but of universal and fundamental values, including the basic right to a fair and impartial trial, and to be allowed to present all the evidence. These are values to which Libya itself subscribes, having signed many international human-rights treaties.

But the Libyan case also involves other values. The first is the humanitarian value of alleviating the tragedy of the infected children. An international fund has been set up to help treat the children in European hospitals, and to strengthen Libya's expertise in dealing with HIV. The international community should continue to strengthen these efforts as part of its solidarity with both the Libyan people and the affected families.

"The protests of the scientific community are a defence of universal values, including the basic right to a fair and impartial trial."

The unfortunate politicization of this case has also diverted attention from another value: the right to safe health care. The transmission of HIV in medical settings in many countries is a large but often 'invisible' problem that is only heard about when it reaches the scale of the Benghazi outbreak, or one in Kazakhstan last summer in which almost 100 children were infected with HIV. There is no internationally recognized set of precautions to make procedures safer, and many nations lack adequate medical supplies and must risk re-using them.

The scientific community, faced with the injustice of the Libyan trial, has acted resolutely. But it must do more to press home the less immediately compelling but equally tragic battlegrounds that the Libyan case highlights in the fight against HIV. ■

Enter *Nature Photonics*

Over the past 50 years, the field of photonics — the scientific study and application of light — has blossomed to become one of the most important enabling technologies of our time. The development of devices such as the laser, the light-emitting diode, the low-loss optical fibre and the CCD (charge-coupled device) detector have transformed the world around us, improving performance in applications such as data communication, materials processing, imaging, biomedicine, lighting and home entertainment.

Fundamental research in photonics is compelling in its own right and promises further transformational technologies. For example, the creation of artificial materials with a negative refractive index is not only cutting-edge multidisciplinary research but holds out the prospect of increasingly high resolution in the detection of, and processing by, light. Engineering materials that trap photons in 'photonic crystals' will, it is hoped, lead to new types of optical memory, and the development of silicon-based all-optical circuits is expected to transform the fields of communications and computation.

Given the excitement and importance of photonics, there is plenty of scope for a journal that captures outstanding research as well as the technologies and their impacts, including commercial developments, and that caters for anyone seriously interested in photonic science and engineering. Accordingly, this month sees the launch of *Nature Photonics* (see www.nature.com/nphoton), whose first issue contains articles spanning the above topics and more.

Nature itself will continue to publish high-impact papers in photonics — the launches of research journals including *Nature Physics*, *Nature Materials* and *Nature Nanotechnology* have never diluted *Nature*'s role in their respective disciplines, but have provided new outlets for top-quality research and discussion.

In publishing terms, *Nature Photonics* is distinctive among the Nature group of journals in being the first to have its editorial headquarters in our offices in Tokyo (supported by associate editors in London and San Francisco). This development will further strengthen editorial links with the research community in Asia Pacific, which is particularly strong in photonics. However, like all Nature journals, *Nature Photonics* is truly international and will publish research from around the globe without geographical preference. ■

RESEARCH HIGHLIGHTS

Food for thought

Biol. Lett. doi:10.1098/rsbl.2006.0566 (2006)

If you stuffed yourself over the festive break, you probably felt the urge to sleep off your heavy meal straight after eating. Grey seals, on the other hand, seem able to postpone digestion until they're ready to rest.

Carol Sparling and her team at the University of St Andrews in Fife, UK, monitored the breathing and heart rate of seals in captivity as they foraged over varying distances. These measurements give the seals' metabolic rate.

The team observed that seals on long expeditions showed peaks in metabolic rate when resting at the surface and overnight, suggesting that this was when digestion happened. The trick may help the seals to juggle the different physiological demands of diving and digestion.

C. SPARLING

OPTICS

Invisibility cloak in sight

Opt. Lett. **32,** 53–55 (2006)

The first material with a negative refractive index for visible light has been constructed.

Light entering a negative-index substance bends in the opposite direction to that entering a conventional material, such as glass. Such materials could be used in new kinds of lenses or even 'invisibility cloaks'.

Metamaterials have already been built to have a negative index for infrared light, with a shortest wavelength of 1,400 nanometres. Gunnar Dolling of the University of Karlsruhe, Germany, and colleagues push down this limit with a metamaterial that works at a wavelength of 780 nanometres, which falls at the red end of the visible spectrum. The researchers built the material by etching an array of holes into layers of silver and magnesium fluoride on a glass substrate.

GENETICS

Pain in the genes

Science **314,** 1930–1933 (2006)

Your sensitivity to pain depends in part on which form you have of a gene that encodes a protein called catechol-*O*-methyltransferase. Luda Diatchenko at the University of North Carolina in Chapel Hill and her colleagues unpick how two common variants of the gene produce differing pain sensitivity, even though they encode the same protein.

The researchers show that RNA transcribed from the variant associated with high pain sensitivity forms a looped structure that inhibits the RNA's translation into protein. This affects pain sensitivity because the protein metabolizes neurotransmitters such as dopamine. Their findings demonstrate one way that a 'silent' genetic variation can affect protein expression.

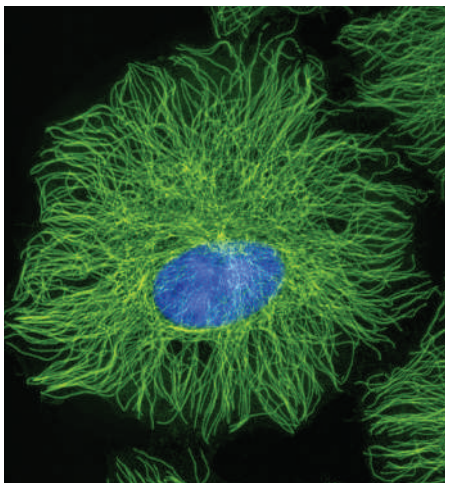
CELL BIOLOGY

Stitched up

Cell doi:10.1016/j.cell.2006.11.025 (2006)

It seems that cells know a few tailoring tricks. Experiments performed by Damian Brunner and Andreas Hoenger of the European Molecular Biology Laboratory in Heidelberg, Germany, and their co-workers suggest that a protein associated with the cell's cytoskeleton acts as both seam tape and a zip.

The researchers studied how a protein from fission yeast known as Mal3p — a homologue of the human EB1 protein — binds to structures called microtubules (pictured below, green). Electron-microscopy images show the protein lined up along the tube's length, following the seam created when the microtubule rolled up from a flat sheet. The researchers show that the protein stabilizes the seam, and speculate that it also helps the tube to curl up, zipping it shut.



T. DEERINCK, NCMIR/SPL

BIOTECHNOLOGY

Meet the VelociMouse

Nature Biotechnol. doi:10.1038/nbt1263 (2006)

Laser surgery can make short work of generating mutant mice, say David Valenzuela and his colleagues at the company Regeneron Pharmaceuticals in Tarrytown, New York.

Mutant mice with deliberately altered genes have become an essential part of the modern scientist's toolkit. But making mice with specific genes 'knocked out' is a laborious process, involving injection of a mouse blastocyst with mutant cells, then several rounds of breeding. The new report describes a short cut: mutant cells are injected into an eight-cell mouse embryo through a perforation opened by a laser. The resulting mice have a greater percentage of mutant cells than those produced by the traditional technique, eliminating the need for breeding. The team dub the method 'VelociMouse'.

NUCLEAR PHYSICS

A double dose of magic

Phys. Rev. Lett. **97,** 242501 (2006)

Unnaturally heavy elements made in particle accelerators tend to fall apart in an instant. Unless they're stabilized by magic, that is.

Jan Dvorak of the Technical University Munich, Germany, and colleagues have confirmed predictions of magic stability for hassium-270, the heaviest 'doubly magic' nucleus seen so far. The 'magic' lies in the

number of protons and neutrons, which are organized into shells like those of electrons in atoms. A filled shell confers relative stability, and hassium-270 is doubly magic because it has filled shells of both protons and neutrons.

The team detected four atoms of hassium-270, created by bombarding a curium target with magnesium ions. They estimate its half-life to be 22 seconds, which is remarkably long for such a superheavy atom.

OPTICS

Two goes into one

Phys. Rev. Lett. **97,** 243003 (2006)

An optical trap uses a highly focused laser beam to confine atoms to a cell known as a potential well. Yevhen Miroshnychenko and his colleagues from the University of Bonn, Germany, have now contrived to squeeze two atoms into a well designed for one.

They achieved this unnatural intimacy through deft manoeuvring of two laser beams: one horizontal, used as a conveyor belt to shunt atoms to and fro; and one vertical, that, like a pair of tweezers, plucks an atom out of one well in the horizontal beam and sets it down in another. The two atoms could form a bound molecule, say the authors, or prove to be a useful resource for quantum information processing.

SEISMOLOGY

Traits of tsunami quakes

Geophys. Res. Lett. **33,** L24308 (2006)

The earthquake that caused a deadly tsunami in Indonesia in July 2006 was a classic 'tsunami earthquake', according to a new analysis of seismic records. Such quakes generate tsunamis much larger than some measures of the quakes' size would predict.

The earthquake near the Java trench sent a wave between 5 and 8 metres tall towards the beaches of Java, killing hundreds. Charles Ammon of the Pennsylvania State University, University Park, and his colleagues show that the rupture that caused it propagated slowly over 170–200 kilometres with an average slip of around 8 metres. The relatively long duration of the event (185 seconds) and the geometry of the slip are features shared with previous tsunami earthquakes.

MOLECULAR BIOLOGY

RNA taken to extremes

Proc. Natl Acad. USA **103,** 19490–19495 (2006)

By comparing the genetic sequences of many bacteria, Ronald Breaker and his colleagues at Yale University in New Haven, Connecticut, have identified a bizarre piece of RNA that they have christened an 'ornate, large, extremophilic RNA'.

The name describes its properties: ornate because the RNA appears to adopt an elaborate three-dimensional structure; large because it is 610 nucleotides long, which is sizeable for a stretch of RNA that does not code for protein; and extremophilic because it crops up mainly in bacterial species that live at extremes of temperature, salt concentration or pH.

The weird RNA may contribute to the membranes of these extreme bacteria and help them to live in such harsh environments. Other similar RNAs could turn up now that people know to look for them.

MICROFLUIDICS

Quick route to crystals

Proc. Natl Acad. Sci. USA **103,** 19243–19248 (2006)

X-ray crystallography can determine the structure of biological molecules, but only if high-quality crystals of the molecule can be grown. Rustem Ismagilov and his co-workers at the University of Chicago, Illinois, present a microfluidics approach to screening crystallization conditions.

C. J. GERDTS

Conventional methods can be time-consuming and require large amounts of the target protein. By contrast, using the new scheme one researcher was able to set up 1,300 crystallization trials in 20 minutes using only 10 microlitres of a protein solution.

The team injected droplets of the solution into a capillary along with another reagent in a way that allowed the reagent and its concentration to be varied in each 'plug' (the picture above shows a 1-metre-long capillary containing around 1,000 plugs). The plugs were then screened to see which conditions gave the best crystals.

JOURNAL CLUB

Pulickel Ajayan
Rensselaer Polytechnic Institute, Troy, New York

Childhood memories cause a nanotechnologist to go nuts for plant-derived nanomaterials.

As a child growing up in Kerala, southern India, I marvelled at the unusual cashew fruit, with its kidney-shaped nut dangling from a swollen apple.

Since then, nanotechnology has become my passion. So it was with a curious mix of scientific interest and childhood memories that I read a recent paper describing how nanomaterials could be derived from plant sources such as the cashew nut.

I had never thought of a cashew nut as anything more than a food item. However, a little research reveals that cashew-nut-shell liquid, rich in natural long-chain phenols, already has applications ranging from hydrophobic coatings to anti-ageing creams.

George John and Praveen Kumar Vemula at the City College of New York, in their recent article (G. John & P. K. Vemula *Soft Matter* **2,** 909–914; 2006), show how cashew-nut-shell liquid can also serve as a starting material for a variety of nanostructures.

The oil contains molecules that have phenol groups for heads, and long hydrocarbon tails. These can form structures such as lipid nanotubes and twisted nanofibres.

To make this happen, the molecules' structure is first modified by attaching water-loving sugar groups to the phenols. The cooperative effect of head groups hydrogen bonding and the hydrophobic interactions of the tails leads the molecules to self assemble into bilayers. These then further organize into the fibres and tubes.

Using a similar strategy, it should be possible to develop a wide range of novel soft nanomaterials from other plant resources. The breadth of precursors available in our plants and crops should inspire all nanotechnologists — not just those fond of cashew nuts.

NEWS

CLONED ANIMALS DEEMED SAFE TO EAT
US regulators prepare to approve food made from cloned animals.
www.nature.com/news

Europe condemns Libyan trial verdict

Bulgaria's accession to the European Union (EU) on 1 January will allow it to apply ever-greater international pressure in the political row over the fate of five Bulgarian nurses and a Palestinian doctor condemned to death in Libya last month.

The six medical workers were sentenced to death on 19 December by the Benghazi Criminal Court for deliberately infecting more than 400 children with HIV at the Al-Fateh Hospital in Benghazi in 1998. Scientists around the world have argued that medical evidence shows unequivocally that the people were not infected deliberately. They point out that the outbreak was a typical example of what can go wrong when hospital equipment and supplies become contaminated — as happened in a hospital in Kazakhstan, where more than 80 children were infected with HIV last summer.

The team defending the medical workers says that it will appeal the verdict to the Supreme Court in Libya. By law, this must be done within 60 days of the verdict. The Supreme Council for Judicial Authority could also annul the death sentences. The council, which makes judicial appointments, is an interface between Libya's supposedly separated executive and judiciary authorities.

Although the strongest criticism of the verdict came from Bulgaria itself, both the EU and Germany, which holds the EU's presidency for the first half of 2007, forcefully condemned the sentences. Bulgaria's new status as an EU member state seems to ensure that this pressure will not slacken.

"We simply cannot accept this verdict," says Benita Ferrero-Waldner, the European Commission's foreign minister. In a letter to the Libyan foreign ministry she pointed to the "recent publication of a strong body of scientific evidence concerning the origin and timing of the Benghazi infection…I very much regret that this new element was not deemed worth considering in the legal proceedings thus far and hope it will be duly taken into consideration by the Supreme Court."

German Chancellor Angela Merkel condemned the verdict as a "terrible ruling"; Frank-Walter Steinmeier, Germany's foreign minister, said that the EU would "continue to exert pressure under the German presidency so that Libya doesn't only take part in a solution but ultimately brings about a solution". This toughened attitude contrasts sharply with that shown by the United States. President George Bush and Secretary of State Condoleezza Rice expressed only "disappointment", and have avoided any discussion of a fair trial or the need for scientific evidence to be taken into account.

The EU's direct language raises the stakes in the power play that surrounds the case. Until now, the international community's approach has mostly been one of 'silent diplomacy' — refraining from public criticism of Libya's handling of the case and relying on behind-the-scenes discussions. These interventions have centred on providing humanitarian aid, which might be seen as compensation (and thus a mitigating factor in Islamic law), while trying not to undercut the medical workers' defence with any implication of guilt and atonement. All the affected Libyan children are being treated in European hospitals.

M. TURKIA/AFP/GETTY

Behind bars: medical workers sentenced to death have attracted global support.

"The EU will continue to exert pressure under the German presidency."

The death sentences mark the failure of this approach, says Emmanual Altit, the French human-rights lawyer who heads the international defence team. Altit has long criticized the politicization of the case, arguing that it acts against the interests of the six medical workers by making them a bargaining chip in Libya's relations with the West.

Now that the political outcry has become noisier, it has been met with anger from Libya. Said Hafyana, the deputy secretary for external relations and international cooperation in Libya's General People's Committee, told Bulgaria's ambassador to Libya that: "No party, no country or authority has the right to intervene in the running of the Libyan justice system or to challenge its equity or fairness." Libyan state-controlled media have also orchestrated a campaign trying to equate the questioning of the guilt of the health workers with indifference to the plight of the children. Some allege that critics are part of a Western conspiracy. "Is the blood of our children mere sewer water?" asked the *El Jamahiriya* newspaper.

When the six medical workers were arrested in 1999, the country's leader Colonel Gaddafi stoked up sentiments such as these by alleging that the infection was a plot by the US Central Intelligence Agency (CIA) and Israel's intelligence agency the Mossad to destabilize the country. Since then, Libya's geopolitical position has changed. After it abandoned the pursuit of weapons of mass destruction in 2003, providing the West with intelligence on nuclear proliferation, the once-rogue state came to be seen as a partner in the 'war on terror'. When diplomatic ties with the United States and the EU were subsequently re-established and trade sanctions lifted, business from America and other nations flocked to the country, which has substantial untapped oil reserves.

The United States is expected to appoint an ambassador to Libya in the coming months, and a visit by Rice seems likely. Sean McCormack, the official spokesman for the US Department of State, says that the verdict itself would not block such moves, and is only one of many considerations. The United States does not intend to take sides in the case, he adds.

Libya's leadership now has greater incentive than ever to avoid diplomatic ructions; but there is substantial resistance to being seen as capitulating to pressure from the West. Gaddafi dug in his heels on 29 December by rejecting calls to release the medical team, and reasserting that intelligence agencies were behind the crime.

According to diplomatic sources, the most optimistic outcome for the moment is that the six medical workers will remain condemned, but that a political solution will be found to have them freed. But the situation is increasingly volatile — and for the moment they remain in grave danger. ■

Declan Butler

See Editorial, page 2.

The dark side of *E. coli*

Last month, the president of fast-food chain Taco Bell appealed to his customers in full-page ads in *The New York Times* and other US newspapers. "You can be confident our food is safe to eat," his letter declared.

The ads were the fallout from a food-poisoning outbreak traced to the chain's restaurants, which has affected at least 70 people across five states. The culprit was *Escherichia coli* O157:H7 — the same as in another outbreak this September and October linked to Californian spinach, which infected nearly 200 across the country and killed three.

The outbreaks have thrown the spotlight on a bacterium that is difficult to detect and virtually impossible to treat or eradicate. "We see it more and more and we don't really know what to do about it," says microbiologist John Fairbrother of the University of Montreal, Canada.

There are thousands of different strains of *E. coli*, most of which are harmless. But O157 can make a potent toxin and latch onto intestinal cells, giving it the ability to cause kidney failure and even death. The bugs live harmlessly in cows' large intestine and are thought to be ubiquitous in cattle lots. Bacteria shed in faeces contaminate meat in slaughterhouses or find their way onto vegetables grown near animals or irrigated with water contaminated with manure, as is thought to have happened with the tainted spinach. Lettuce is thought to have spread the Taco Bell outbreak.

There are now some promising research leads that might help prevent future outbreaks. At a meeting earlier this year on pathogenic *E. coli*, veterinary researcher David Smith of the University of Nebraska, Lincoln, and his colleagues reported that a vaccine containing proteins from O157 cut the number of cows shedding bacteria by 60–70%. Canadian company Bioniche Life Sciences, based in Belleville, Ontario, has submitted the vaccine for regulatory approval in Canada, and plans to do so in the United States.

Other groups are turning to viruses, called bacteriophages, that attack the O157 strain. A group led by microbiologist Todd Callaway of the US Department of Agriculture's Food and Feed Safety Research Unit in College Station, Texas, has found that feeding sheep a mixture of bacteriophages cuts the number of pathogenic bacteria in their guts by over 1,000 times.

Cattle farmers may be forced to adopt vaccines or therapies because of pressure from food processors and the threat of lawsuits. But some microbiologists question whether these

Green menace? *E. coli* O157 can spread if greens such as spinach are irrigated with tainted water.

J. ERNST/REUTERS

European funding targets big biology

Metagenomics and the human variome project can expect substantial boosts in the European Union's next round of science funding, starting this month.

Unlike its predecessors, which covered five-year periods, Framework Programme 7 (FP7) runs for seven years, until 2013. It has a budget of more than €50 billion (US$66 billion), of which two-thirds is earmarked for large public–private collaborations in ten areas, including health, transport and nanoscience.

The first calls for proposals went out on 22 December. For the health theme, the focus is on big biology. A planned metagenomics project, one large topic for the first call, aims to study the vast community of microbes living in the human body, which are thought to influence physiology, nutrition and immunity.

"There are some two kilograms of microorganisms living in the human body," says Dusko Ehrlich, a microbial geneticist at the French agricultural research agency INRA in Jouy-en-Josas. Most of them can't be grown in the lab: "We just don't know what's in there." Proposals are requested for projects to sequence the microbial genes present in the body, which are thought to outnumber our own genes by a factor of 100.

Another project earmarked for funding will support the global human variome project, which was launched in Melbourne, Australia, last June. The variome is the set of variations between different people's genomes, which influence the development of disease and drug side-effects. As yet there is no global system for collecting and sharing information on the human variome; FP7 promises to support one.

FP7 also provides €7.5 billion for a new Europe-wide funding agency called the European Research Council (ERC), which launches this month. The council will fund smaller projects that do not meet the industrial or societal goals required by the rest of the framework programme. ■

Nora Eichinger

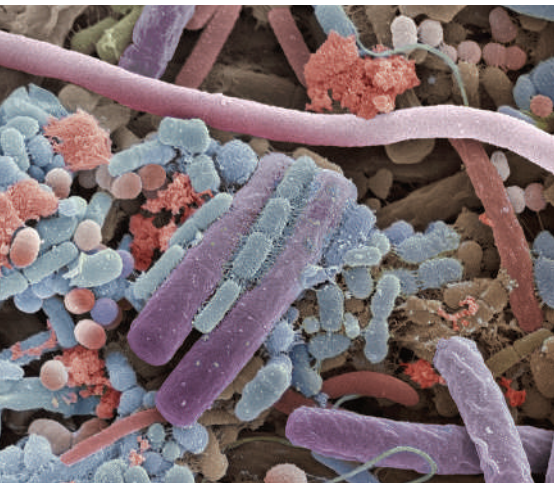
reductions will protect humans. Just ten cells of O157 are enough to infect a person, compared with hundreds of thousands needed for a *Salmonella* or cholera infection. "It's a very different standard for foods to meet," says James Kaper, an *E. coli* expert at the University of Maryland, Baltimore. He notes that irradiation would rid food of the bacteria but that the public, food industry and food-safety regulators have been reluctant to adopt it.

So researchers are also working on treatments, for example with antibodies that inactivate the toxin. (Antibiotics aren't recommended for *E. coli* because by the time the infection is diagnosed, the bugs have usually released so much toxin that killing them doesn't help.) But researchers admit that the demand for such drugs is likely to be low.

Perhaps an underestimated problem are other pathogenic *E. coli* strains, including O26, O111 and O145. These can also cause serious food poisoning but are more likely to go unnoticed, because lab tests are more difficult or not routine. These strains are more common outside the United States; in Italy, for example, most cases of *E. coli* food poisoning probably go undetected, says Alfredo Caprioli, who directs the *E. coli* reference lab at the Istituto Superiore di Sanità in Rome. There is intense interest in finding the exact combination of genes that make strains harmful to humans and quick ways to test for them.

Ultimately researchers must find the critical points in the food supply at which intervention can most reduce contamination, says food scientist Don Schaffner of Rutgers University, New Jersey. "Obviously we haven't studied it enough to solve the problem." ■

Helen Pearson



S. GSCHMEISSNER/SPL

Fellow travellers: the European Union will fund studies of humans' vast microbial flora.

PLANS FOR 2007?
Share your thoughts and plans for the new year.
http://blogs.nature.com/news/blog/2006/12/plans_for_2007.html

Open-access journal will publish first, judge later

A radical project from the Public Library of Science (PLoS), the most prominent publisher in the open-access movement, is setting out to challenge academia's obsession with journal status and impact factors.

The online-only ***PLoS One***, which launched on 20 December, will publish any paper that is methodologically sound. Supporters say the approach will remove some of the inefficiencies associated with current peer-review systems — but critics question whether a journal that eschews impact factors will manage to attract papers.

Among the 90 or so papers in ***PLoS One*** at its launch are reports on the meaning of wild gibbon songs and a mathematical model of rabies control. The authors of both papers say they chose ***PLoS One*** because they support open access, and because they wanted to be part of something new. "I think we're seeing one of the future directions of scientific publishing," says Colin Russell, a zoologist at the University of Cambridge, UK, and an author of the rabies paper.

"We're trying to make a journal where papers are not the end point, they are the start of a discussion."

Every paper submitted to the journal is reviewed by at least one member of ***PLoS One***'s editorial board of over 200 researchers, but only to check for serious flaws in the way the experiment was conducted and analysed. In contrast to almost all other journals, referees ignore the significance of the result. Notable papers will instead be highlighted by the attention they attract after publication.

Visitors to the ***PLoS One*** website can, for example, attach comments to specific parts of a paper and rate the paper as a whole. Data from those systems, as well as download and citation statistics, will then allow ***PLoS One***'s editors to identify and promote the papers that researchers are talking about. "We're trying to make a journal where papers are not the end point, they are the start of a discussion," says ***PLoS One*** managing editor Chris Surridge, based in Cambridge, UK.

The journal will initially publish 10–15 papers a week, but this could reach a few hundred each month, says Surridge. The system makes sense, he says, because a single review process avoids the time wasted when papers are rejected from high-ranking journals and reviewed again elsewhere. Others add that the journal's decision to accept papers from all areas of science could benefit authors of interdisciplinary studies, whose work is often rejected by subject-specific journals.

But ***PLoS One*** faces some significant challenges. Many new journals struggle to attract papers until they are given an impact factor (a measure of the citations its papers receive), but a journal that accepts everything can't usefully be classified in this way. Critics also point out that referees may be reluctant to review potentially trivial papers, and that existing journals have had little luck persuading readers to comment on papers after publication.

Yet Surridge is bullish about his journal's chances. He thinks referees will appreciate the approach, as it will cut the number of reviews that scientists as a whole have to make. He adds that existing attempts to encourage comments don't reflect the way scientists actually read papers — something he aims to remedy by allowing visitors to highlight and annotate different sections of text. Surridge also says that other systems offered little reward to researchers; ***PLoS One*** will allow comments to be rated by others, letting users establish status accordingly.

Rival publishers have suggested that ***PLoS One*** is an attempt to prop up PLoS's finances (see *Nature* 441, 914; 2006). At present, PLoS relies on annual philanthropic donations of several million dollars to break even. The only similar open-access publishing venture — the online-only journals run by BioMed Central — is only now close to breaking even, six years after launch. But Surridge shrugs off the criticism, saying that ***PLoS One*** is designed to meet PLoS's aim of making scientific literature freely available. ■

Jim Giles

SPECIAL REPORT

Alien Earth

With improved techniques, growing data sets and a new space mission, 2007 is the first year in which we might discover another planet like our own. **Katharine Sanderson** reports.

When marvelling at the stars on a clear night, it's hard to imagine that there are up to 400 billion of them in our Galaxy alone. Even harder to comprehend is how many planets may be orbiting these stars — a number that could run into trillions. Surely somewhere among them there must be a comfortable home for alien life, even if it's not advanced enough to be gazing back at us?

This is the question that exoplanet hunters are trying to answer. So far, they have spotted 209 planets beyond our Solar System. These tend to be gas giants in searingly hot orbits close to their parent stars — unlikely to be habitable. But researchers are edging closer to finding the one type of planet that we know can support life — a carbon copy of our own Earth. Thanks to improved techniques, mounting data and a new space mission, many believe that 2007 could be the year we find the first truly Earth-like planet. At the very least, we should have a much better idea of how common alien Earths may be.

The main obstacle for planet hunters is that planets outside the Solar System are obscured by the light from their stars, so our telescopes can't see them directly. Most researchers make use of the fact that when a planet orbits a star, its gravitational pull causes the star to wobble slightly. As the star wobbles, its speed as seen from Earth (its radial velocity, or RV) changes, and this shows as a change in the wavelength of the star's light. This can be used to estimate a lower limit for the planet's mass.

Unfortunately, RV tends to detect big planets that are close to their stars. Heavier planets cause more obvious wobbles. And close planets have shorter orbits, allowing researchers to observe several wobbles over a relatively short time.

So far, the method has found 197 exoplanets, the smallest of which is at least 7.4 times the mass of Earth. But improvements in accuracy are allowing researchers to spot ever smaller planets, and as more observations are made, it's possible to detect planets that are farther from their stars. Geoffrey Marcy of the University of California, Berkeley, is responsible for an impressive 121 of the planets found using RV, and says his team is set to announce several exciting discoveries in 2007. He adds that he expects RV to reveal several rocky planets this year, and that within the next few years we may know whether Earth-like planets are common or rare in the Universe.

Transit tricks

Another limitation of RV is that, on its own, it doesn't reveal what a planet is made of. But coupling its mass with an estimate of the planet's radius — which can be worked out as it passes in front of (transits) its star — gives the density, which indicates whether a planet is a diffuse gas giant or a smaller rock. The main problem with this is that to see a transiting planet you need to look at the correct angle at the correct time, and with most exoplanets found we haven't been so lucky. Transits are most common for large planets close to their stars, which again skews our knowledge away from planets like Earth.

In 2007, however, a European mission called COROT aims to improve the odds. Launched on 27 December, it will look specifically for transiting planets, which will then be followed up with RV measurements, either from Earth or from NASA's Spitzer Space Telescope.

> "COROT will find oodles of short-period planets, both hot Jupiters and hot super-Earths."

After a couple of months of calibration, COROT will make observations for 150 days, then turn 180° and observe for another 150 days. To obtain reliable data, scientists need to watch three or four transits, so the craft's short observing period limits the search to planets close to their stars. This means COROT is unlikely to find a true Earth analogue. But it will tell us a lot about how common large, rocky planets are. "It's a necessary proof," says Malcolm Fridlund, project scientist on COROT. Fridlund is proud that Europe is leading the field rather than NASA. "We're taking the first step for once," he says.

"If rocky planets are common enough, COROT could find planets that are habitable," says David Charbonneau, an astronomer at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.

The first results from COROT should be in and analysed by mid-2007, with ground-based follow-up experiments done by the end of the year. "COROT will find oodles of short-period planets, both hot Jupiters and hot super-Earths," says Alan Boss, a theoretician at the Carnegie Institution in Washington DC. But some see COROT simply as a stats-gathering prelude to NASA's more ambitious Kepler mission, which is due to launch late in 2008 and will make observations over four years. "Kepler holds even greater promise to detect rocky planets around Sun-like stars," says Marcy.

For 2007, the best chance of spotting an Earth-like planet may come from another ground-based technique: gravitational microlensing. When two stars are closely aligned along our line of sight, the front star acts as a lens and magnifies light from the star behind, sometimes enough for an orbiting planet to be seen directly. Whereas RV is most sensitive to planets close to their stars, microlensing works best for distantly orbiting planets. Four planets have so far been detected in this way, including an icy 'super-Earth' (J.-P. Beaulieu *et al. Nature* **439,** 437–440; 2006).

Not alone? The search for Earth-like planets beyond our Solar System is hotting up.

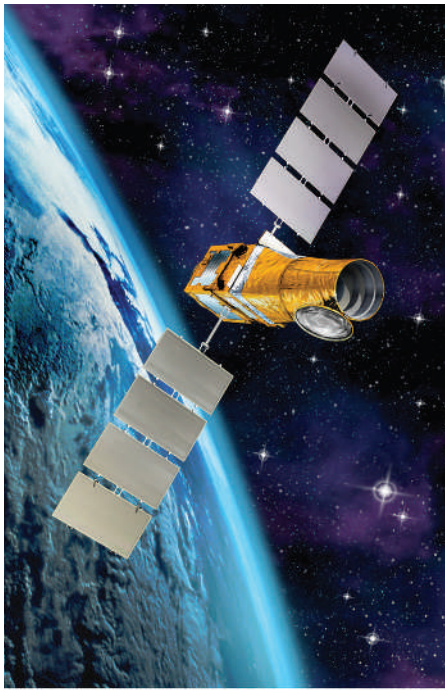
MARS IN FOCUS
Find all our news on the red planet online.
www.nature.com/news/infocus/mars.html

NASA

A star undergoes a microlensing event as seen from Earth only once every 100,000 years, so spotting them is "pretty hairy" says Andrew Gould of Ohio State University in Columbus. Gould coordinates a team of astronomers around the globe (the MicroFUN group) who monitor changes in the brightness of around 50 million stars. They post a list of possible microlensing sites online, and Gould alerts astronomers in the appropriate locations where and when to look.

The technique has bagged the lowest-mass planet so far, at 5.5 Earth masses. Gould predicts a one in ten chance of spotting an Earth-size planet in 2007 using microlensing, and reckons that in five years' time, he may be able to spot ten per year.

If an alien Earth is found, the big question of whether it could support life will require information about its atmosphere. And researchers are set to make big progress here too this year. It's possible to scan a distant planet's atmosphere by taking the spectrum of the light emitted by its star when the planet is completely hidden from view, and subtracting it from the combined spectrum when the planet is in front. The technique was used to investigate the spectra of Pluto and Charon in the 1980s, and in 2001, Charbonneau and his team used the Hubble Space Telescope to detect sodium in the atmosphere of a Jupiter-class planet orbiting the star HD 209458.



D.DUCROS/CNES

The COROT spacecraft will spend most of this year looking for rocky planets that might host life.

Last month, Jeremy Richardson of NASA's Goddard Space Flight Center in Greenbelt, Maryland, and his colleagues reported at the American Geophysical Union meeting in San Francisco that they have now used the Spitzer Space Telescope to make a more detailed study of the infrared spectrum of the same exoplanet. They measured the radiation emitted by the planet at various wavelengths, and are hopeful that they will be able to glean some information about the composition of its atmosphere. The team doesn't want to reveal details ahead of publication early this year. But co-author Sara Seager says that although these first spectral measurements are "pretty crummy", the work is crucial proof that the technique will "provide the first-ever spectrum of an exoplanet. This is a big breakthrough in exoplanet science."

This will be a big year for Spitzer, adds Charbonneau. He predicts that the telescope will study ten planets in detail in 2007, making temperature measurements and detecting gases such as methane and carbon monoxide.

For the future

Looking beyond 2007, "Kepler is the big one", says Charbonneau. Kepler's goal is to find an Earth-like planet. Like COROT, it will seek out transits, but will have more time to spot planets farther out from their stars.

After that, prospects for the field are less rosy. NASA recently cut funding for its proposed Space Interferometry Mission (SIM), which is now seriously delayed. SIM would make very precise measurements of stars' wobbles in two planes, which would give an actual mass, rather than the lower limits provided by RV.

The decision to cut funding has not gone down well, with Gould citing "stupidity" as one reason for the cut. Boss says the situation is "depressing" and a backward step for exoplanetary research. He warns that even a slight dip in funding could cause well-trained scientists to leave the field. Two other projects, the European Space Agency's Darwin and NASA's Terrestrial Planet Finder, would take the first images of Earth-like planets. But these are also receiving too little funding to proceed.

Should planet hunters be allocated scarce funds ahead of other projects? Scott Gaudi, who works with Gould, argues that the search for other worlds is a societal imperative: the question of whether other Earths exist is one that humans have always asked, he says. Despite the funding problems, he remains upbeat: "There is a very good chance that in our lifetime we will answer the age-old question: is there life out there?" ■

Japanese universities fire researchers for misconduct

Two leading Japanese research universities have taken the rare step of firing faculty members who had been the subject of misconduct investigations.

On 20 December, Osaka University dismissed biologist Akio Sugino, three months after its investigation concluded that he fabricated data in two papers on DNA replication (W. Nakai *et al. J. Biol. Chem.* doi:10.1074/jbc.M603586200; 2006 and T. Seki *et al. J. Biol. Chem.* 281, 21422–21432; 2006). Sugino has admitted that he faked the data for the first paper, which was withdrawn in August, but has refused to retract the second.

On 27 December, the University of Tokyo fired biochemist Kazunari Taira and his lab researcher Hiroaki Kawasaki, saying that they had "damaged the university's honour and credit". A university investigation concluded in April 2006 that "there was no reproducibility and no credibility" in four papers on RNA from Taira's lab, although the researchers deny misconduct (see *Nature* 440, 720–721; 2006). It is the first time one of the university's faculty members has been dismissed for research-related problems.

Project BioShield loses supplier for vaccine

The US government has cancelled the biggest and most visible contract under its $5.6-billion Project BioShield programme, telling a small California company that it failed to make agreed progress in developing an anthrax vaccine for civilian biodefence.

The cancellation last month leaves the government without a major supplier of next-generation vaccine. Michael Leavitt, the government's secretary of health and human services, said he would make finding one "a priority".

The company, VaxGen, signed an $878-million contract in 2004 to develop 75 million doses of a recombinant anthrax vaccine. But it has faced chronic problems, and in May said it would deliver the vaccine at least two years late (see *Nature* 441, 281; 2006). The company, which has the right to appeal the termination, says that it is "actively exploring its strategic and legal alternatives".

B. SKLAR/AP

VaxGen has lost its lucrative contract to make anthrax vaccines for the US government.

Media coverage of womb pictures had fetal flaw

Last November, an amazing picture of an elephant inside its mother's womb was carried in newspapers and magazines around the world (see *Nature* **444,** 529; 2006). But what our report and many others failed to mention was that the picture was actually of a silicon model, created with information from ultrasonography.

The images are part of a National Geographic documentary called *In the Womb*. A press release issued by a television channel led to the media reporting several model pictures as ultrasound images of real fetuses.

To correct the error, *Nature* is happy to publish one of the original ultrasound images used to create the models. Just as breathtaking, it was taken by Thomas Hildebrandt of the Institute for Zoo and Wildlife Research in Berlin, Germany, and a team at the African Lion Safari in Cambridge, Ontario. This elephant is 125 mm long and weighs just 200 grams, and will continue to gestate for another 16.5 months.

T. B. HILDEBRANDT

Creationists back down over schoolbook stickers

A school board in Cobb County, Georgia, has dropped its fight to get warning labels placed on high-school textbooks that teach evolution.

The stickers, which state that evolution is "a theory, not a fact", were originally approved by the board in 2002, after some parents complained about a new biology textbook. A second group of parents, with the help of the American Civil Liberties Union, sued over the stickers, and in 2005 a district judge ruled that they should be removed.

The school board had a right to appeal but decided to abandon the case, in part because of rising costs. The 19 December announcement came almost a year to the day after a federal judge ruled that intelligent design — the concept that an external designer shaped evolution — could not be taught in Dover, Pennsylvania (see *Nature* 439, 6; 2006).

School boards in Texas, Oklahoma and Nevada are mulling over legislation to 'teach the controversy' over evolution in 2007.

Subsidy cuts secure rise for Japan's research budget

The overall budget for science and technology in Japan will decline by 1.8% to ¥3.5 trillion (US$29 billion) for fiscal year 2007, the government announced on 27 December. But the main research budget will increase by 1.1% to ¥1.3 trillion, as cutbacks will focus on administrative expenses and subsidies to universities and research institutes.

Large-scale national projects that began last year, such as advanced supercomputers and an X-ray free-electron laser, will see a substantial increase. Smaller projects — even those in the eight strategic research fields such as the environment — will not see much gain.

The country's life-sciences budget will edge up by 0.5% to ¥68.8 billion, with the big winners including translational research and proteomics. But funding for a project in personalized medicine, including the establishment of a BioBank, will fall by 16% to ¥2.6 billion.

Europe's astronomy club gains a thirteenth member

The Czech Republic — where 400 years ago Johannes Kepler famously deduced that planets move in elliptical orbits around the Sun — has joined the European Southern Observatory (ESO). That makes it the astronomy agency's first member from central or eastern Europe.

On 22 December, Czech education minister Miroslava Kopicová signed an agreement that, when ratified by the country's parliament, will make the Czech Republic a full member. It will be the thirteenth member of the ESO, which was formed in 1962 with five member countries in an attempt to strengthen Europe's research in astronomy and astrophysics.

The ESO operates a number of telescopes at three facilities in the high desert of Chile, including the four 8.2-metre telescopes that comprise the Very Large Telescope.

BUSINESS

When the party's over

A drug-trial failure leaves Pfizer in search of a new corporate strategy to deal with the post-blockbuster age, as **Meredith Wadman** reports.

The world's largest drug company starts 2007 in need of a fresh start. Most of all, Pfizer has to put aside the end of 2006, when it was forced to pull the plug on its eagerly anticipated cholesterol-lowering drug torcetrapib. The compound was found to be associated with unacceptably high death rates in a late-stage clinical trial involving 15,000 people (see *Nature* **444,** 794–795; 2006).

This is no run-of-the-mill failure. Pfizer's current cholesterol drug Lipitor brings the company more than $12 billion in annual revenues, but loses patent protection in 2011. The company was counting on torcetrapib to replace those sales. The news sent Pfizer shares down 10% in a day, dissolving $21 billion in market capitalization.

There's still an outside chance that Pfizer could replace Lipitor: two chemical cousins of torcetrapib — known as cholesterol esterase transfer protein (CETP) inhibitors — are in early development. Steven Nissen at the Cleveland Clinic is using ultrasound scanning of torcetrapib's effect on plaque build-up in the coronary arteries to see if its toxicity was a quirk — in which case the other compounds might prove viable.

But no one is betting on it. For Pfizer's 106,000 staff, the failed trial was the culmination of an inauspicious year. Long-time chief executive Hank McKinnell was replaced in July by lawyer Jeffrey Kindler. And only days before the drug trial was abandoned on 2 December, the company said it would lay off more than 2,000 sales representatives — 20% of the sales force. This was seen as another move in Pfizer's quest to keep Wall Street satisfied in the face of expiring patents and anaemic drug pipelines.

In this environment the company's 13,000 researchers are entitled to worry about their future. "There appears to be a pattern to right-sizing the organization, that probably implies either trimming research and development or, at least, investigating whether it should be trimmed," says Tony Butler, a pharmaceuticals analyst with Lehman Brothers in New York.

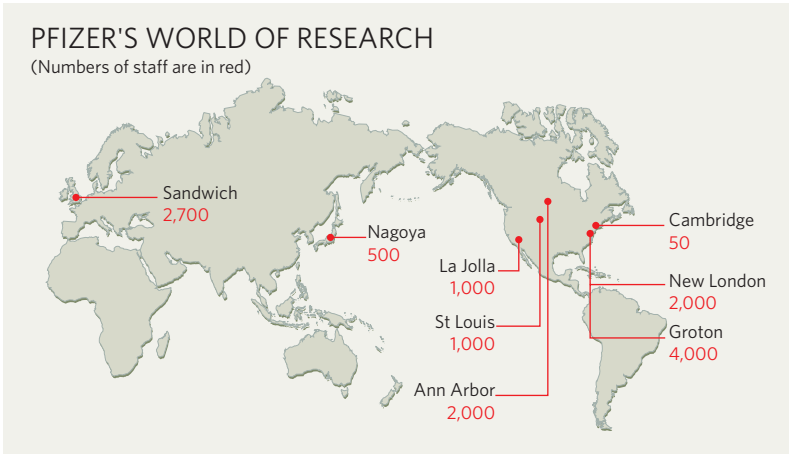
Peter Rost, a company gadfly and former Pfizer marketing vice-president, now in litigation with the firm over the circumstances of his departure in 2005, is more direct. "It's very likely that Pfizer is going to pull back on personnel in all areas, including research," he says. Rost's blog, http://peterrost.blogspot.com, has been abuzz with chat on the circumstances and implications of the trial failure.

F. BECKHAM/AP

Jeffrey Kindler: Pfizer's new chief executive has pledged to cut costs aggressively.

The company hasn't publicly discussed its strategic position in the wake of December's setback and declined to comment for this article. But in the press release that announced the trial failure, Kindler pledged to lower costs "as expeditiously as possible". He indicated that research and development, along with manufacturing and other branches of the business, could be subject to cost-cutting.

Pfizer's research and development operation cost $7.4 billion in 2005 and comprises eight main laboratories (see map). It researches drug discovery in 11 major fields, from heart disease to cancer.



The company has already downsized its Pharmaceutical Sciences division, which helps decide which drug candidates to take forward into clinical trials, and invents the process for manufacturing them. Four years ago, this division had a budget of $1 billion and 3,800 employees; today it is down to 2,700 people, with an operating budget of $800 million.

Pfizer has also embraced the industry-wide trend towards acquiring or partnering with smaller companies — and in some cases big ones — to bolster its drug pipeline. But apart from its 2000 acquisition of Warner-Lambert, which gave it Lipitor, this strategy has provided few obvious successes.

Niche products

Analysts say that putting its own laboratories on the chopping block won't help Pfizer unless it recognizes that the era of blockbuster drugs — those generating more than $1 billion in annual revenue — is over. According to Boston-based management consultants Bain & Company, among many others, Pfizer and the rest of the pharmaceutical industry need to develop more sophisticated drugs, targeted at smaller numbers of people.

Pfizer currently has 242 products at various stages in its pipeline, including plenty that might fit that bill. They include compounds that harness the immune system to fight cancer, and an obesity drug similar to Sanofi's rimonabant (see *Nature* **437,** 618–619; 2005). Butler thinks that concerns about the pipeline have been overblown: "It looks pretty good to me," he says.

On Wall Street, however, a drug that promises sales of 'only' $500 million won't generate much excitement in the shares of a company with annual sales of over $50 billion.

"Pfizer was one of the few remaining companies that appeared to be able to manage the mega-blockbuster drugs," explains Kenneth Kaitin, the director of the Tufts Center for the Study of Drug Development in Medford, Massachusetts. "But what this whole episode demonstrates is that Pfizer is not immune to the enormous risk and costliness of drug development — and the problems that come with late failures," he says. "It must adapt its research and development strategy to focus on products that address smaller markets — and that can be brought to market more quickly, efficiently and at a lower cost." ■

That's oil, folks...

Optimists see oil gushing for decades; pessimists see the planet's energy future already drying up. **Alexandra Witze** reports.

Don't say they didn't warn us. The poster for the meeting of the Association for the Study of Peak Oil and Gas in Boston this October featured American revolutionary Paul Revere on his midnight ride, bringing news of imminent calamity. Only this time it is not the British who are coming, but the end of the oil era, and with it much of western civilization. Many attendees at the meeting were people who could tell you how to stock a bunker to survive the inevitable collapse of civilization, and then opine at length about the extent and characteristics of the great tar-sand deposits of Canada. Some of them conduct a thriving mini-business in preparing for the coming apocalypse — "deal with reality or reality will deal with you", as one website claims — while scrutinizing table after table of data on world oil production.

But this is not an easily dismissed fringe. Respected geologists with lifetimes of experience are genuinely concerned that the world is about to see an unprecedented crisis — a reduction in the supply of a primary fuel before an alternative is available. When we moved from wood to coal, it was not for a shortage of forests; when we moved in large part from coal to oil and gas, it wasn't because the pits were empty. But many people are convinced that the flow of oil is destined to start falling, and soon.

Matthew Simmons, an energy investment banker in Houston, Texas, and self-described "petro-pessimist", argues that the world's great oilfields are moving quickly towards the end of their production, or have already passed into rapid decline. The North Sea, for instance, is the only place that a significant new discovery has been made outside of nations in the Organization of the Petroleum Exporting Countries (OPEC), Russia and Alaska in the past four decades. It is now in eclipse — production in the region peaked in 1999, which is earlier, Simmons says, than expected. The United Kingdom no longer exports oil, he notes, and production in Norway — the North Sea's long-term stalwart — is also declining. And no new giant oilfields are taking the place of those that have already passed their peaks, says Simmons.

> **"I don't know if we are all Hubbertists now, but we are all recognizing that there is a finite quantity of oil."**
> — Robert Kaufmann

Some people think that the declines we are seeing are indicators that the world is on the verge of, or has already passed, the maximum amount of oil that can physically be produced. In their view, oil production follows a bell-shaped trajectory, with the peak occurring when half of the total reserves have been consumed. Therefore, it should, in theory, be easy to determine whether the peak has already occurred or whether it is yet to come. Total up the world's oil reserves, estimate the rate at which countries have produced oil, and you'll know where you are in the trajectory. If the reserves are more or less equal to the amount already pumped, then production is at its peak.

If you accept this principle, then the issue of when the peak comes depends mainly on the amount of reserves that remain untapped, and that in itself gives room for disagreement. But some don't accept these premises. To them, these arguments are simplistic geological determinism that does not take into account the role of oil prices. To the dissenters, reserves are not a geological given but a function of the current price and the extraction technology that price can buy. New reserves will be developed as the market demands.

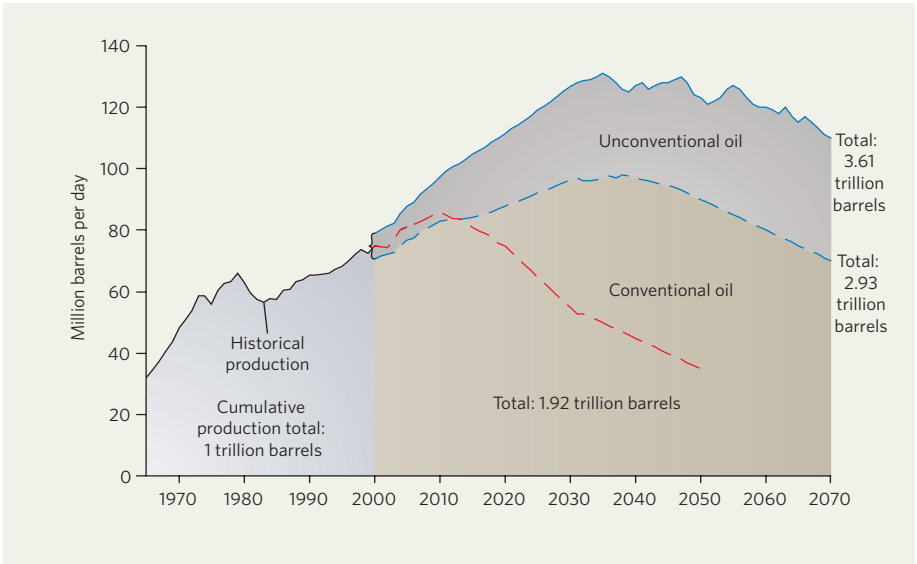
Opposites attract

Both sides issue regular, well-referenced reports that come to opposite conclusions on whether the world is running out of oil. "There's just no middle ground," says Kenneth Deffeyes, a geologist who has retired from Princeton University in New Jersey, and is a leading supporter of the peak-oil theory. His personal belief is that we are already a year past the peak. If he's wrong, though, he's sure that it will prove not to be by much: the peak is imminent, and unavoidable.

Meanwhile, a study from energy analysts Cambridge Energy Research Associates (CERA) in Massachusetts sees no sign of any peak in production occurring before 2030. And, crucially, CERA doesn't see a peak with a steep downside — rather a crest followed by an undulating plateau (see chart), which would be much less apocalyptic even if it happened today.

It's not that CERA thinks that oil production has no constraints, or that the geological resource can't be depleted. Even oil company executives say publicly that they see a problem. T. Boone Pickens, the maverick Texas oil magnate, has said that he thinks oil production may already have reached its maximum. And in October, during an address to the National Press Club in Washington DC, Shell Oil president John Hofmeister acknowledged that "the easy stuff is running out". "We may argue about when the peak is, but it doesn't change the argument that it's coming," says Robert Kaufmann, an energy economist at Boston University in Massachusetts. "I don't know if we are all Hubbertists now, but we are all recognizing that there is a finite quantity of oil."

SOURCE: CERA



Twin peaks: peak-oil supporters think we have already reached or will soon reach a historical maximum of oil production (red line); others argue that oil production will not peak until at least 2030 (blue lines).

J. KAPUSTA

Hubbert, in this context, is M. King Hubbert, the geophysicist who first predicted that oil production would peak quite suddenly — and that when it did, it would slump sharply thereafter. In 1956, while working in Shell Oil's research laboratory in Houston, Texas, Hubbert predicted[1] that oil production in the contiguous 48 states of the United States would peak in the early 1970s. Hubbert's calculations produce a bell curve to describe the rate of oil production, with a sharp rise on one side of the peak and a symmetrical drop-off on the other (see chart, overleaf).

At the time Hubbert made these calculations less than half of this two-sided curve had been seen. Oil exploration and discovery were booming, and Hubbert's prediction looked implausibly pessimistic. But he turned out to be right; production in the contiguous United States reached its peak in 1970, and almost overnight Hubbert gained his own personal fan club.

Those in favour of the peak-oil theory argue that Hubbert's methods for analyzing US oil output can also be used to analyze the global production peak. Deffeyes, known to many as the charismatic protagonist of *Basin and Range*, the first of John McPhee's great popular accounts of modern geology, has emerged as a particularly prominent Hubbertist. In *Hubbert's Peak: The Impending Oil Shortage*[2] and *Beyond Oil: The View from Hubbert's Peak*[3] he used the same mathematics as Hubbert to calculate the total oil reserves worldwide that remain to be produced. With a flourish of exactitude, Deffeyes estimated that the world reached peak-oil production on 24 November 2005. He later pushed this back — but only to 16 December of that year.

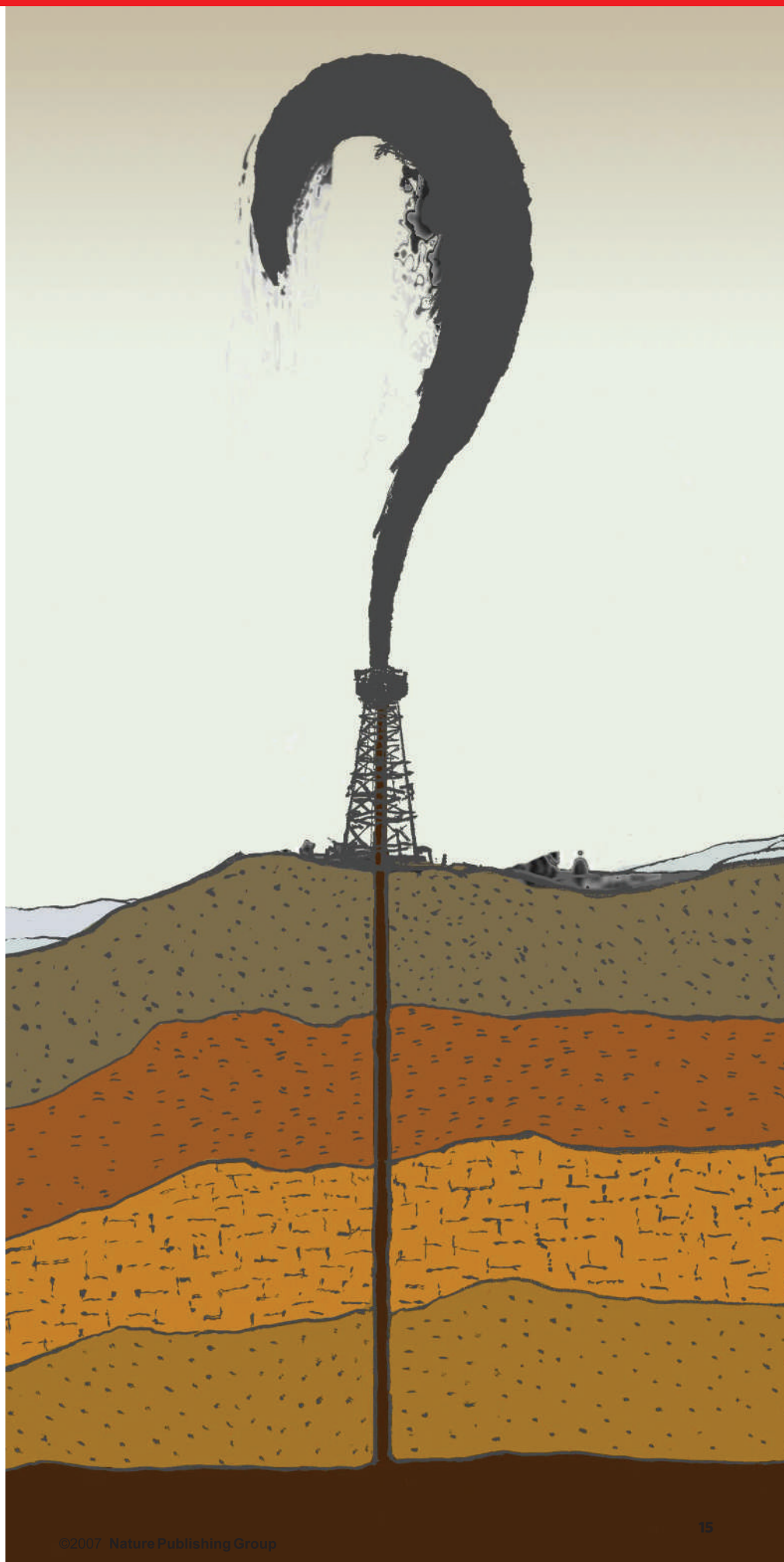
"I'm not declaring victory just yet," he says hastily. He's saving the victory announcement until he sees world production numbers drop for three years in a row. "If I'm right, it will be very transparent in five years."

Changing with the times

But not everyone buys Deffeyes' interpretations. "The technique that Hubbert used in the 1950s is simply not applicable on a global basis in 2006," argues Peter Jackson, an oil and gas analyst who works for CERA near London, UK. Jackson authored CERA's background briefing paper "Why The Peak Oil Theory Falls Down" in November 2006.

Jackson notes that proponents of the peak-oil theory have changed their dates several times before; as production numbers come in year after year, for instance, leading peak-oil theorist Colin Campbell has postponed his estimates for the peak in all hydrocarbon production from around the turn of the millennium to 2010. Others point out that predictions of an unavoidable slump are almost as old as the oil business; John Strong Newberry, chief geologist of the state of Ohio, was predicting that America's oil would soon be tapped out back in 1875.

Jackson, Deffeyes and everyone else in the debate agree that nearly 1.1 trillion barrels (175 trillion litres) of oil have been produced



C. AURNESS/CORBIS

Pump it up: the amount of oil recoverable depends on our capacity to access it.

worldwide. A key difference is that supporters of the peak-oil theory argue that roughly that same amount remains to be pumped out of the earth's reserves, whereas CERA's report estimates those reserves at 3.7 trillion barrels, a number that would place the world well on this side of any peak in production.

One reason for the difference is that CERA is much more optimistic about the amount of oil that can be recovered from operating oilfields through the use of new technologies. Jackson points out that the expected recovery estimates for operating oilfields often grow with time. The total estimate of recoverable oil from the North Slope of Alaska, for instance, used to be 9.6 billion barrels; today it is 13.7 billion barrels.

"If I were to say we are not finding enough oil every year through exploration to replace what we are producing, you would be alarmed," says Jackson. "And that is correct. But peak-oil supporters don't talk about field reserve upgrades — in a lot of the producing fields around the world, companies are constantly updating estimates of reservoir reserves."

Reserves change in size even after the initial geological mapping of an oilfield because the amount of oil that's recoverable — and thus deserves to be counted as reserves — depends on the skill of the oil companies and the effort they are willing to put in. Globally, about 35% of the oil present in established fields is actually produced. Nearly every oilfield matures through the same sequence: first, the easily recovered oil is extracted through traditional drilling. When that runs low, engineers begin a process of secondary extraction, using techniques such as injecting water or carbon dioxide to drive more oil out of the rock. Many fields also undergo tertiary extraction to squeeze out yet more oil, usually by injecting steam to lower the viscosity of the oil. Oil wells are abandoned once the cost of extraction is no longer worth it. "We will still have oil in 100 million years," says David Hughes, a geologist with the Geological Survey of Canada in Calgary. "It just won't be recoverable at an energy profit."

But the fluctuating price of oil means that fields abandoned after secondary production can be re-opened for tertiary production when demand calls. The current high price of oil — just over US$60 a barrel — provides an incentive for companies to start tapping into their reserves and pushing into more areas for discovery. In general, those against the peak-oil theory claim that the Hubbert-curve approach underestimates changes in extraction technology brought about by both natural developments and changes in the price of oil, which can turn fields that are too costly to pump from into valid reserves.

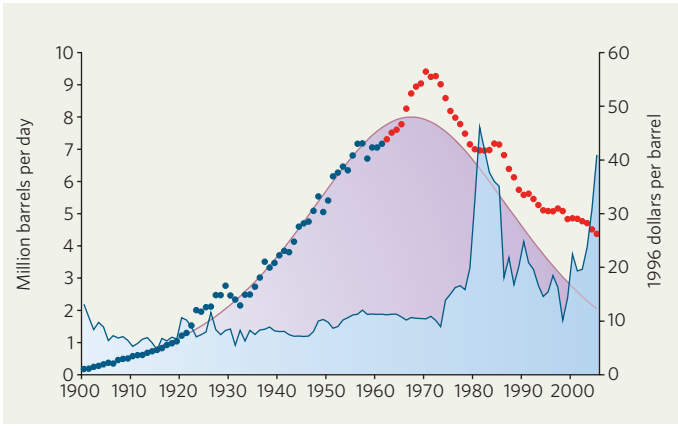
As the great oilfields of the world age — most of them are now undergoing secondary, if not tertiary, extraction — other discoveries could help to plug the supply gap, argues Jackson. In 2000, an analysis by the US Geological Survey of petroleum reserves estimated that there were 1 trillion more barrels of oil worldwide than previously thought. The survey estimated that any worldwide peak in oil production wouldn't happen until 2030 at the earliest. In part as a result of the high prices of the past few years, roughly 500 oil-development projects are slated to start producing oil in the next five years, says Jackson, and these will range in size from an estimated million barrels per day down to 10,000 barrels per day.

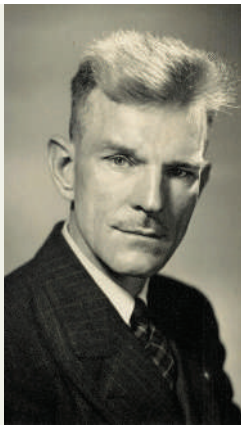
In the deep

Energy optimists point to recent discoveries such as the seven new oilfields uncovered in the deep waters of the Gulf of Mexico last year — reserves that are only accessible because of new technology. The biggest oilfield found there to date, the Thunder Horse field, is estimated to contain 300 million barrels of oil. Such exploration may also soon be helped by the US Congress, which voted on its last day of session in 2006 to approve expanded drilling into previously off-limits areas in the Gulf of Mexico. "There's a lot of new capacity coming to the market," says Jackson. "That's one of the reasons I'm not too concerned about a peak."

Another reason for doubt is that Hubbert's model was not perfect even when applied just to the contiguous 48 states of the United States — its great claim to fame. Although the date predicted for the peak was roughly correct, the model predicted an amount of production at the peak that was 20% less than reality. And, in part because of unforeseen discoveries in the

SOURCE: R. KAUFMANN





WWW.MKINGHUBBERT.COM

Fan club: M. King Hubbert (right) gained instant notoriety for his 1956 prediction that oil production in the 48 contiguous US states (purple shaded area) would peak around 1970. It did, but production was much higher that year and in later years (red dots) than Hubbert foresaw.

Gulf of Mexico, the amount of oil produced in the United States after the peak was much greater than Hubbert had predicted.

But the peak-oil theorists are not convinced. "The problem is, if you go and talk to people whose job it is to actually go and find this stuff, they have no clue as to where these trillion barrels of reserves actually are," says Michael Rodgers of the energy analysis firm PFC Energy in Washington DC.

Who to trust?

Estimates of reserves that are published by oil companies, national governments and researchers such as the US Geological Survey are not to be trusted, according to the peak-oil supporters. Jeremy Gilbert, former chief petroleum engineer for BP, says that not all oil companies work to the same standards. The US Securities and Exchange Commission sets rules for how to report reservoir estimates, but only US and major international companies generally abide by those standards — and they don't always do so reliably. "The standards for other national companies are unknown," Gilbert says. "If someone tells you the reserves in Kuwait are 75 billion barrels, he has no idea how that 75 billion was calculated." Peak-oil supporters are eager to point out that after a sharp drop in the oil price in the mid-1980s the estimated reserves of various OPEC countries — including Iran and Iraq, which had a mutual war to finance at the time — were jacked up by their governments.

Partly as a result of such manoeuvres, Simmons is particularly pessimistic about whether OPEC nations can continue to slake the world's thirst for oil. Currently, OPEC countries provide about 40% of the world's oil. Non-OPEC countries have been consistently producing more oil than they've been finding since the late 1980s, Rodgers told the meeting in Boston. He thinks that production from non-OPEC countries will peak between 2010 and 2015; after that, no amount of OPEC production can make up the gap between supply and the world's growing demand. The fact that such predictions have been made before does not necessarily mean that they are wrong now.

Most of the attention on OPEC focuses on Saudi Arabia, by far the biggest producer and the driver of oil prices worldwide. The country gets most of its oil from seven giant maturing oilfields. The three biggest fields have been producing oil for more than 50 years, and the oil industry constantly swirls with rumours that the biggest of all — the Ghawar field — has been increasingly cut with water to drive its production even higher. Simmons, after studying technical reports published by the Society of Petroleum Engineers, has argued that the state-run oil company Saudi Aramco routinely overestimates the country's oil reserves. Saudi Arabia, he argues, is closer to running out of oil than most people think.

The 1970s oil crisis, when OPEC slapped an embargo on countries that had supported Israel, was only a taste of things to come, says Rodgers. At the time, the United States was the only country that had already peaked in oil production. Now, many more countries — including Peru, Argentina, Norway, Congo and Mexico — have also passed their peak. Countries such as Canada, China, Brunei and Malaysia are currently undulating around a plateau of oil production and could soon decline, he predicts.

New discoveries, such as in the Mexican section of the Gulf of Mexico or in Angola or Brazil, could change the date of an imminent oil peak, says Rodgers, but only slightly. "All you can do is take the cliff facing us in the next few years, and push it farther out over time. This has already happened." Fields of the size now being discovered, though, will not be large enough to push the peak back far. The 300 million barrels at Thunder Horse provide less than a week's consumption at the world's current rate of use.

Another way Russia helped out in the 1990s, Simmons points out, was by producing less oil than had been expected. That apparently helpful drop, though, was an exception. "[Misstated reserves] wouldn't be so bad if demand remained steady," says Simmons, "but instead it soared." The world produced 16% more oil in 2005 than it did in 1990 — and none of that production, Simmons argues, came from any major new discovery. Instead, it came from a compilation of much more incremental discoveries and increased production from a number of countries. Meanwhile, demand is expected to continue to grow as more and more families buy cars worldwide.

> **"All you can do is take the cliff facing us in the next few years, and push it farther out over time."**
> — Michael Rodgers

A dipstick for the future

Some of the oil production to meet this future demand may come from alternative approaches — extracting oil from oil shale, for example, or liquefying natural gas or coal for fuel (see *Nature* **444,** 677–678; 2006). These 'unconventional' sources of oil are some of the reasons that the CERA outlook is so optimistic. But many of the other sources, say peak-oil supporters, are not good alternatives — certainly not the sort of thing that can easily be used in vast quantities as a replacement for sweet Saudi crude, a high-quality oil with a low sulphur content, even if the price is right. In Alberta, Canada, geologists have focused on extracting oil from tar sands which could, in theory, help supply the world's needs for decades. But the process is expensive, both financially and environmentally; three barrels of water, for instance, are needed to produce each barrel of oil from the sands, and the production releases large amounts of greenhouse gases.

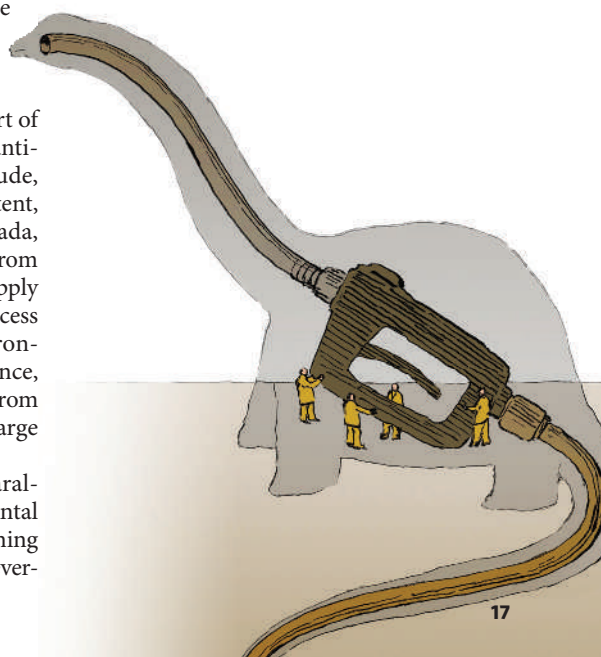
Thinking along those lines raises a parallel question: can we afford, in environmental terms, to put the peak off, and to keep turning oil into atmospheric carbon dioxide at an ever-increasing rate? From an environmental point of view, a peak might almost be welcome. If the subsequent rapid drop in production crashed the world economy, though — in the way that peak-oil supporters fear — those benefits might be hard to appreciate. What's more, the resources needed to develop the alternatives on which economic recovery would depend might not be available.

Those problems might be lessened if the peak could somehow be predicted. But Kaufmann, the economist, says not to expect any financial or other indicators. Oil prices didn't rise sharply or otherwise indicate an imminent depletion of US oil resources just before the peak in 1970, he says, mainly because the cost of production was staying stable. To predict peak oil in advance, "you need some kind of nice price signal," he says, "and we don't see any of those signs yet."

One place to look for such signals might conceivably be in the prices for which oil is bought and sold on the futures market. And at the moment, the New York mercantile exchange is settling on prices around US$67 a barrel. It's a price high enough to make alternative fuels interesting, but in real terms not remarkably high compared with long-term averages. If oil production does start to collapse, peak-oil supporters who want to stock their bunkers with luxury goods have the opportunity to make a killing, by buying tomorrow's oil comparatively cheap and selling it, when the time comes, much more dearly. If, that is, the time does actually come. ■

Alexandra Witze is *Nature*'s Chief of Correspondents, America.

1. Hubbert, M. K. Shell Development Company Publication 95, June 1956.
2. Deffeyes, K. *Hubbert's Peak: The Impending Oil Shortage* (Princeton University Press, 2006).
3. Deffeyes, K. *Beyond Oil: The View from Hubbert's Peak* (Farrar Straus Giroux, 2006).



Life's a game

Manipulating society has traditionally been the preserve of politicians and the gods. Does the current boom in virtual worlds give social scientists and economists an opportunity to join them? **Jim Giles** investigates.

SONY ONLINE ENTERTAINMENT

Is a ruthless dictatorship a better way of running a country than a well-oiled democracy? Would people be happier if all their property was confiscated? Might our economies be healthier if inflation ran at 100%?

We're sure we know the answers. Kim Jong-il's reign in North Korea is a brutal one. Hyperinflation is wrecking Zimbabwe. History tells us that stealing land prompts wars, not happiness. By comparing these and other examples, we can be confident that each proposal is, well, foolish.

But how would an experimental scientist view this kind of knowledge? Many physicists answer questions by running experiments, tweaking conditions and recording the results. They do so repeatedly, until the role of every variable is understood. Case studies are the beginning of the process, not the end. And physical laws are not limited to observations of what approach seems to work best; they can produce complex models that allow predictions of the future.

The difference in approach is no reflection on the abilities of social scientists. It's just that societies are not so amenable to experimentation. Government economists would not be popular if they repeatedly tweaked monetary policy just to see what happened. Social scientists learn from history. They run surveys. They even conduct small experiments with a handful of subjects in idealized conditions. What they can't do is manipulate the system they are studying.

Enter Edward Castronova, an economist at Indiana University in Bloomington, occasional writer of fiction and an expert on multiplayer online games. His vision is nothing if not ambitious: to create societies with the intention of experimenting on them. The societies will exist only inside computers, with real people living some of their lives through characters (avatars) in these virtual worlds. But Castronova does not intend to merely simulate real life. He and other researchers want to tweak the rules of those worlds so they can study everything from democracy to monetary policy. Such tools, says Castronova, would be the "supercolliders" for his field. They would usher in an era of "computational social science".

Parallel worlds

If that sounds unlikely, visit Azeroth. You'll have plenty of company: more than seven million people play World of Warcraft, the online role-playing game that is based in this virtual realm. The game, run by Blizzard Entertainment of Irvine, California, is pure axe-wielding tolkienian fantasy: players go on quests and slay mythical beasts, amassing virtual wealth and power in the process. Nevertheless, Azeroth society shares many features with our own. Players produce goods and trade them; they cooperate to achieve goals but also conduct personal vendettas; some are unhelpful and rude, others — admittedly not many — display altruism.

Elsewhere in cyberspace, economies are springing up that, superficially at least, seem to mirror the real world. Cyber-capitalism is most apparent in the malls of Second Life, an online world that is visited by around 10,000 people a day. The world's designers, Linden Lab of San Francisco, have eschewed fantasy quests and given users the tools and virtual land they need to build their own online experience. Since Second Life's launch in 2003, universities have leased space to teach classes, and high-street retailers have set up outlets where avatars can get their hair cut or buy virtual roller-skates. The in-world currency now floats against the dollar on dedicated exchanges, and one user claimed last November that her Second Life property business has made her a real-world millionaire.

Virtual worlds also show similarities to real life at the level of one-on-one social interactions (despite many players choosing outlandish avatars such as giant mice). In a paper in press with *CyberPsychology & Behavior*, Nick

Yee at Stanford University, California, shows that some of the unwritten rules of real-life socializing appear to cross over into Second Life, even though communication is purely text based. Male avatars stand further apart than females when talking, for instance, and tend to make less eye contact.

If these cyber-societies follow some of the rules of real life, what can social scientists do with them? Many want to treat them like a biologist treats cells in a Petri dish. Tinker with monetary policy. Rewrite the rules of democracy. Force players to work together, or try and drive them apart. But conduct each intervention in a systematic way, holding all other variables constant and rigorously monitoring the outcomes. Thanks to the huge audiences that online games attract, social scientists will then be manipulating society itself. Although a full "supercollider" experiment has not yet been done, tentative steps have been taken, and the results suggest that the idea has merit.

VIVENDI/BLIZZARD ENTERTAINMENT

Multiplayer games like World of Warcraft attract millions of online players.

Location, location

First up was a study of a seemingly trivial question: why are certain businesses based in specific places? The answer is also apparently trivial: they are there because we say they are. It makes sense for actors and film producers to agree that movies are made in Hollywood and Mumbai. Anyone who wants to be in the film business then knows where to go. There is nothing that compels movie stars to behave in this way, except that everyone involved agrees on the location — what social scientists call a 'coordination effect'.

At least that's one theory. One can also argue that there is something special about Hollywood or Mumbai that means that movies get made there, for example, and that coordination effects are less important. It has been hard to settle the argument either way, which undermines attempts by economists to explain how societies reach such agreements. That then hampers our understanding of bigger questions, such as how cooperative behaviour evolved to become such a prominent feature of everyday life.

This uncertainty persists in part because social scientists have only limited ways of studying coordination effects in real life. Distinguishing between rival explanations is tricky because we can't re-run history and, for example, probe the factors that drew movie moguls to California in the first place.

Except that now we can, because some virtual worlds regularly recreate themselves to cope with overcrowding. Take Norrath, the land in the fantasy game EverQuest. After launching in 1999, EverQuest's blend of trolls and spells proved so popular that Sony Online Entertainment, which runs the game, created additional copies of Norrath on multiple servers. There are currently 26 servers running. And each time Sony creates another world, a new society evolves under almost exactly identical starting conditions.

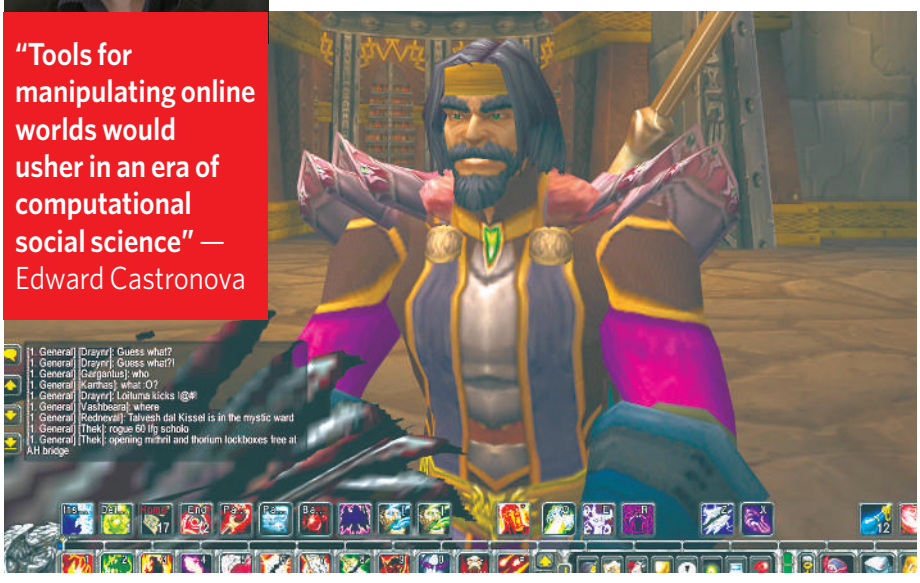
This cloning of worlds gave Castronova his first chance to do computational social science. Using a survey of EverQuest players (E. Castronova *Games and Culture* **1,** 163–186; 2006), he showed that on each server just one region has become established as a market. Crucially, that region differs between servers, although mountain ranges and cities have identical locations in all the worlds. So there does not seem to be a single prime location for the market; instead, some chance event seeds its creation, and coordination effects then lock it into place. "With no small amount of trepidation," Castronova writes in a footnote to the paper, "I would venture to claim that this is the first time in human history that a distinct macro-social phenomenon has actually been verified experimentally."

For Castronova, this result hints at what could be achieved. But other social scientists are more sceptical. When *Nature* contacted authors of some highly cited papers on coordination effects, they expressed interest in Castronova's approach, but none had yet read his paper. After doing so, they highlighted various shortcomings they saw in the study.

Some took issue with the paper's more grandiose statements. Castronova's "trepidation" is well justified, says Andrew Colman, a game theorist at the University of Leicester, UK. Colman's laboratory studies of coordination effects, along with other work, had indicated their importance before Castronova's study. And Colman's studies involve more than 80

Castronova has his own online avatar in World of Warcraft (below) and is now creating a world of his own. world.

"Tools for manipulating online worlds would usher in an era of computational social science" — Edward Castronova



VIVENDI/BLIZZARD ENT; C. MEYER/INDIANA UNIV.

students — around the same number of players from each server that responded to Castronova's survey. Other researchers questioned whether the worlds that Castronova studied are truly independent, as it would be possible for players to communicate outside the game.

Castronova acknowledges that lab work has already shown that coordination effects are real, and says that although his study needed just a few survey respondents to identify the market location, the coordination effect resulted from the interaction of hundreds of thousands of people. "It's a dramatic difference in terms of scale and complexity." Researchers might, for example, be interested in how coordination effects are influenced by information campaigns run by governments — an issue that is beyond the scope of lab work.

But the biggest criticism of this approach is one that all virtual-world experiments will have to face: "The number-one challenge is generalizability to the real world," says Dmitri Williams, a social scientist at the University of Illinois at Urbana-Champaign, who is credited along with Castronova for developing the idea of using virtual worlds to study real society.

To illustrate the problem, Williams points to a 'plague' that broke out in World of Warcraft in summer 2005. At first, says Williams, researchers proposed that the spread of the disease could be used as a model for real-world epidemiology. But it soon became clear that some players were deliberately infecting others, because, unlike in real life, they suffered little penalty for doing so. Players are also much more willing to risk death online than they are in the real world. "The risks don't match up," says Williams, "so we have to be very careful about generalizing our conclusions."

Nevertheless, Williams is keen to use virtual worlds to examine theories about how communities form. Like many of those working on virtual worlds, he feels that data from these experiments complement rather than replace those from traditional social-science studies.

Castronova agrees, adding that watching mice in mazes does not reveal everything about how the human brain works, but it does give researchers another way of studying the issue. For now, the only 'mice' that Castronova and others can study are those playing games such as EverQuest. The designers have often welcomed researchers who want to study cyber-societies, but this partnership can go only so far. Game companies cannot be expected to mess with their profitable products to please social scientists, so to realize his goal, Castronova needs to build his own world to rule over.

"The number-one challenge is generalizability to the real world. The risks don't match up"
— Dmitri Williams

That world, or at least a prototype, should be here soon. Using a one-year grant of $250,000 from the John D. and Catherine T. MacArthur Foundation, based in Chicago, Illinois, Castronova and his colleagues are building the Shakespeare-themed world of Arden. Fifteenth-century maps of the English county of Somerset will define a landscape over which blacksmiths, tavern-keepers and bards will roam. A test version should be up and running by March, and its results might enable Castronova to attract the funding needed to build a fully functional version. Ultimately, he wants around 500 people to play for 100 hours per month each — enough to provide a functioning economy. The players will be there because it's fun.

To run economic experiments of interest to Castronova, Arden will need to develop an economy that features aspects of the real world, such as inflation. By building different monetary conditions into different versions of Arden, Castronova will be able to test aspects of economic theory such as supply and demand. With two versions of Arden with different prices for a particular good, theory says that demand should be higher in the world where the good costs less. This is just an example, as Castronova will not reveal exactly what experiment he is planning for fear of invalidating the study.

Arden could in principle be used to test any idea that interests social scientists. One might examine how laws influence individual behaviour, suggests law researcher Dan Hunter of the University of Pennsylvania in Philadelphia. Others want to test new forms of democratic participation or assumptions in marketing theory.

Reality check

Enthusiasm is no promise of success, however. Perhaps the most immediate problem facing Castronova's team, if they secure longer-term funding, will be game design. Sony's multimillion-dollar budget buys teams of experienced game designers. Castronova has around 30 committed but less qualified members of his university. He must also balance the need to design a world that he can control for his research with the need to create an experience that is enjoyable enough to get people coming back. No university team has built a large-scale online game for research purposes before, and there is no guarantee of success.

Castronova will also have to grapple with the sometimes unanticipated ethical problems that arise in online research. In one study of the playing habits of online gamers, a group of US-based social scientists collected anonymized data from a game company. But researchers who were regular game players soon realized they had enough information to link avatars to people they knew.

Other problems may emerge as games become more popular. There is real money at stake in many games, and players have been known to contact the police over thefts of cyber-weapons. Greg Lastowka, at Rutgers School of Law in Camden, New Jersey, says that players may increasingly take real-world legal action to remedy online crimes, undermining the chances of testing new legal systems in the worlds themselves.

Computational social science is yet to prove itself as an established research method. Right now, admits Castronova, many social scientists dismiss such studies because they feature fantastical avatars in mythical quests. And to prove them wrong, solid results are needed. But as befits a man on a mission, Castronova is confident that he can provide them: "They'll believe us when we start showing it's possible". ■

Jim Giles is a reporter for *Nature* based in London.

LINDEN RESEARCH

Does shopping for virtual goods imitate shopping habits in real life?

RIKEN aids international structural genomics efforts

SIR — We would like to respond to comments made in your News story "'Big science' protein project under fire" (*Nature* **443,** 382; 2006) about Japan's Protein 3000 Project. This government project funds nine centres, including the RIKEN Structural Genomics/Proteomics Initiative, consisting of the RIKEN Genomic Sciences Center's Protein Research Group (http://protein.gsc.riken.jp) and the RIKEN SPring-8 Center.

First, we do not agree that the information gained is "of limited use" or, as one researcher is quoted as saying "a lot of it is junk". RIKEN has made major contributions to structures and structural models of functionally important proteins. It is expected to have determined 2,500 new structures by spring 2007, or 5% of the entire Protein Data Bank (PDB). Nearly half of those determined so far were obtained using NMR, and consist of functional domains from biologically important (including disease-related, signal transduction and nucleic-acid-binding) human/mouse proteins. Multiple structures from each family were analysed to understand binding specificities. About 70% of all NMR structures of human/mouse proteins deposited in the PDB in 2005 were from RIKEN.

The Protein 3000 project contributes significantly to the goals of the International Structural Genomics Organization (www.isgo.org) by providing large numbers of templates that can be used to model other members of the protein families. On average, each NMR structure from RIKEN has contributed to about 300 new homology models, and each X-ray structure to about 200 new models at a level of 30% sequence identity. Quality assessment measures for RIKEN structures are similar to those for structures deposited in the PDB in 2000–2006 by traditional structural biology groups, according to Gaetano Montelione of the Northeast Structural Genomics Consortium (personal communication).

Second, we very strongly disagree with the comment that "A centre of that size should contribute to methodology, but there has been nothing." RIKEN has made seminal contributions to the development of methodologies and technologies. RIKEN has pioneered cell-free protein synthesis on a production scale, and developed technologies for extensive sample optimization process using the cell-free method. These technologies have been indispensable in solving the structures of many difficult proteins. More than 1,000 NMR structures have been determined from protein samples synthesized by RIKEN's implementation of the cell-free method. Additionally, at RIKEN, N. Kobayashi has developed the KUJIRA software for spectral analysis and P. Güntert has developed the CYANA software for automated protein structure analysis. RIKEN will be opening the NMR facility, together with these important technologies, to external scientists in 2007.

Shigeyuki Yokoyama*, Thomas C. Terwilliger†
*RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan
†Los Alamos National Laboratory;
on behalf of the ISGO Executive Committee
This letter is also signed by:
Seiki Kuramitsu, *RIKEN SPring-8 Center, Japan*
Dino Moras, *Institut de Génétique et de Biologie Moléculaire et Cellulaire, France*
Joel L. Sussman, *Israel Structural Proteomics Center, Israel*

Advances in biology reveal truth about prokaryotes

SIR — Although we agree with William Martin and Eugene V. Koonin's point in Correspondence ("A positive definition of prokaryotes" *Nature* **442,** 868; 2006) about the validity of the term 'prokaryote', a term that Norman R. Pace has proposed abolishing ("Time for a change" *Nature* **441,** 289; 2006), they have lost sight of the organismic biology forest for the molecular biology trees. The main differences between prokaryotic and eukaryotic cells probably relate to the original symbioses from which eukaryotes evolved.

Eukaryotes — whether protoctists, fungi, animals or plants — routinely open their membranes to take in (or let out) nuclear genomes, whole cells or other large particles, in processes such as ingestion, fertilization and hybridization. They reseal their membranes and live happily ever after. All eukaryotic sexuality requires cell fusion. Nearly all eukaryotic cell phenomena involve microscopically visible intracellular motility that never happens in prokaryotes.

We need to reassess our understanding of the course of evolution by recognition of the differences between unidirectional transfer of genetic material as the basis of prokaryotic sexuality — genophore DNA, viruses, plasmids — and parental cell fusion in eukaryotes. Roger Stanier and Cornelius van Niel's concept of 'prokaryote' was brilliantly recognized in 1927 by Boris Kozo-Polyansky, who only wrote in Russian. The word 'procariotique' was independently coined in 1925 by Edouard Chatton for cyanobacteria and all other bacteria including archaebacteria (J. Sapp *International Microbiology* **9,** 163–172; 2006). Because of the modern developments of biochemistry and molecular biology, electron microscopy and comparative genetics, the term 'prokaryote' is even more valid now than it was when first introduced.

Michael F. Dolan, Lynn Margulis
Department of Geosciences, University of Massachusetts–Amherst, 611 North Pleasant St, Amherst, Massachusetts 01003-9297, USA

Pollution analysis flawed by statistical model

SIR — I found your Special Report on air-pollution control in the United States ("The politics of breathing" *Nature* **444,** 248–249; 2006) to be generally well balanced. I would like to point out, however, that there is by no means universal agreement among scientists that air pollution at contemporary US levels affects human health. I am one of the sceptics.

The report seems to take at face value the conclusion of "two large, well-respected epidemiological studies", that every additional microgram of fine particles per cubic metre in the air causes tens of thousands of deaths a year in the United States. Yet joint pollutant analyses — with sulphur dioxide and either sulphates or fine particles both included in the statistical models — show that sulphur dioxide is associated with mortality; fine particles are not (D. Krewski *et al. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality* Health Effects Institute, 2002). The association of sulphur dioxide with mortality remains unexplained, as there is no plausible biological mechanism by which it could be causing death.

Further, the pollution studies mentioned used the proportional hazards model for analyses of the data. This model assumes that the relative risks of air pollution and potential confounders remain constant over time. It is clear, however, that the basic assumption of proportionality of hazards is satisfied neither for air pollution nor for a strong potential confounder, cigarette smoking (S. H. Moolgavkar *Inhal. Toxicol.* **18,** 93–94; 2006). Use of this model when the assumption of proportionality of hazards is violated can have serious consequences for the inferences drawn from the data. It may, for example, explain the very different results of observational epidemiological studies of hormone replacement therapy in the 1990s and the recently concluded Women's Health Initiative randomized trial (R. L. Prentice *et al. Am. J. Epidemiol.* **162,** 404–414; 2005). Departing from assumptions of proportionality of hazards for potential confounders may also bias the estimates of main effects in cohort studies, particularly when the confounder is a strong risk factor. In air-pollution studies, the use of a manifestly wrong model to adjust for confounding by smoking probably biases the estimates of small air-pollution effects on mortality, although the direction of the bias will depend upon the structure of the correlation between smoking and air pollution.

We do not currently have the methods to reliably estimate small environmental risks.

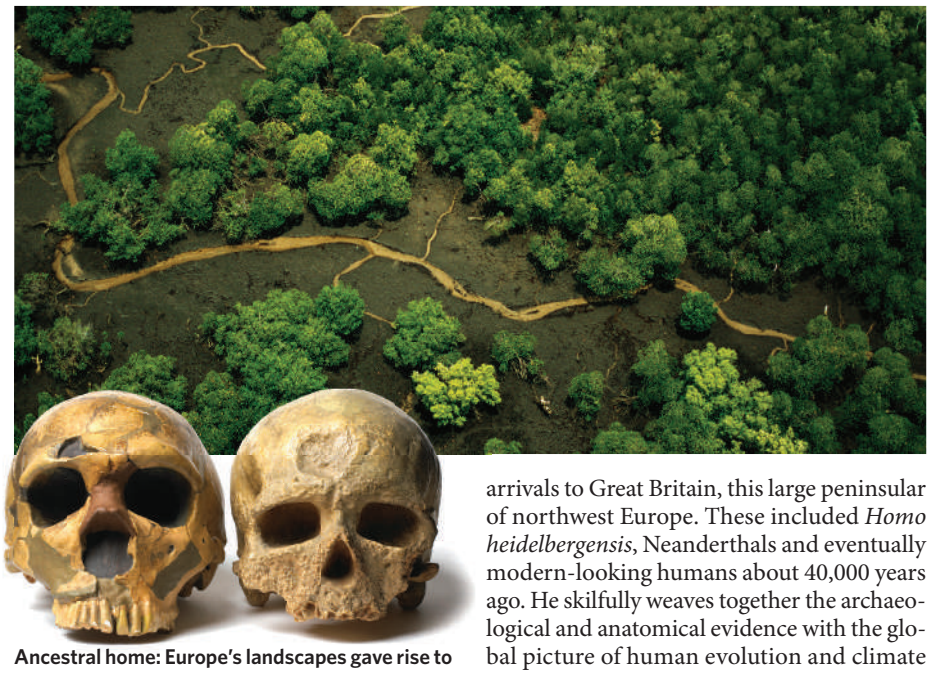
Suresh Moolgavkar
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, Washington 98109, USA

BOOKS & ARTS

The making of Britain

Who are the "amphibious ill-born mob" who gave rise to the British nation?

Homo Britannicus: The Incredible Story of Human Life in Britain
by Chris Stringer
Allen Lane: 2006. 320 pp. £25
The Origins of the British: A Genetic Detective Story
by Stephen Oppenheimer
Constable & Robinson: 2006. 534 pp. £20
Blood of the Isles: Exploring the Genetic Roots of our Tribal History
by Bryan Sykes
Bantam Dell: 2006. 400 pp. £17.99. Published in the US as *Saxons, Vikings, and Celts*. W. W. Norton: 320 pp. $26.95



B. HAAS/GETTY IMAGES

B. AUDUREAU/NATURAL HISTORY MUSEUM

Ancestral home: Europe's landscapes gave rise to Neanderthal (left) and Cro-Magnon man.

Clive Gamble

Who do the British think they are? The nature and future of Britishness has always been high on the political agenda. Who better than Daniel Defoe, author of *Robinson Crusoe*, to characterize their island condition; Celts, Romans, Saxons, Normans and Vikings, "From the amphibious ill-born mob began/ That vain ill-natured thing, an Englishman". Not surprisingly, Britain's chancellor of the exchequer, Gordon Brown, sees positive values for the nation's post-imperial legacy in the "long tidal flows of history — from the 2,000 years of successive waves of invasion, immigration, assimilation and trading partnerships". In a speech earlier this year he argued that the distinctive set of values these flows produced will enable Britain to seize the opportunities of globalization because the nation is "stable, outward-looking, committed to scientific progress and the value of education".

National character is important but its cultural and biological roots are notoriously open to political manipulation. The question is, where beneath the rallying flags of nationalism, patriotism and ethnicity can we find common ground for science and education to meet without undue controversy? Certainly not in the cultural field, where the movement of peoples is concerned. Without a whiff of irony for an American living in London, the poet T. S. Eliot declared in 1948 that "it would appear to be for the best that the great majority of human beings should go on living in the place where they were born". Eliot's point was that cultural strength came from local tradition. He went on to say that it was fine when whole prehistoric tribes moved and so transposed their cultures wholesale. It was the piecemeal movement of individuals and cultures that offended him, and this continues to resound in today's debates on immigration and asylum seekers.

Eliot and Defoe would have taken little comfort from these three books on the earliest inhabitants of Britain. But I hope that Brown will, although their conclusions have significant policy implications for Britishness that I will return to later.

Homo Britannicus takes us furthest back in time. In an excellent and engaging account of the Ancient Human Occupation of Britain Project, which he masterminded and leads, Chris Stringer of the Natural History Museum in London starts 800,000 years ago, when Britain was neither an island nor inhabited. The Thames was a mere trickle, and the major river, the now vanished Bytham, flowed out to sea north of present-day Ipswich. It was along that coast at Happisburgh and Pakefield that the project's researchers, helped by local residents, discovered and excavated some stone tools. Their discoveries added almost 300,000 years to the history of Britain.

Then came the massive Anglian glaciation that buried the Bytham and pushed the drainage south to the Thames. With the aid of beautiful illustrations, Stringer charts the many arrivals to Great Britain, this large peninsular of northwest Europe. These included *Homo heidelbergensis*, Neanderthals and eventually modern-looking humans about 40,000 years ago. He skilfully weaves together the archaeological and anatomical evidence with the global picture of human evolution and climate change. The book is a triumph, communicating the trials and thrills of scientific research across many disciplines to provide answers to the coevolution of environment and people.

But are the ancestors under the *Homo Britannicus* umbrella that common ground where science and education can meet? Stringer provides the basal strata, and the authors of the two other books under review use genetics to reconsider the varied contributions of Defoe's "amphibious ill-born mob". *The Origins of the British* by Stephen Oppenheimer is particularly illuminating. The author carefully lays out the genetic data that show how three-quarters of Britishness dates to the repopulation of the peninsula 15,000 years ago, after the northern ice sheets last retreated. This was long before agriculture, and millennia before rising sea levels made us into 'little Britain'. These influential ancestors were hunters and gatherers, and the genetic data trace their movement from Iberia along a seaboard route to western Britain; there was also a smaller influx into England from northern Europe. Once he has established this fundamental east–west divide, Oppenheimer takes us through a fascinating investigation of what this means for some cherished notions of Britishness. He shows that Old English was indigenous, not imported by the Saxons, and shows how unimportant the Vikings are to the story. He is kinder to the Celts, providing

evidence that the recent debunking of them as a nationalist myth has gone too far. His chief input to the Britishness debate is that biologically and culturally, the Anglo-Saxons were not the first English nation.

Bryan Sykes agrees. In *Blood of the Isles*, a shorter and less well-illustrated volume supported by a website, he reports on the Oxford Genetic Atlas Project. The added interest is his account of racist scientists such as Robert Knox, who distinguished between the hard-working Anglo-Saxons and the indolent Celts on the basis of hair colour and head size. Sykes is particularly good at demolishing these repugnant myths with his genetic data, although I was alarmed by his impudent claim that "my art is oblivious to the prejudice of the human mind". Historical genetics is just as much an interpretation for its time as the shape of skulls was in the nineteenth century.

So, where does all this leave Britishness? As Sykes says, "this really is the history of the people, by the people". We carry our past in our genes and, as Oppenheimer shows, if we are looking for that common ground where science and education meet, then it was 15,000 years ago when small groups of highly mobile hunters entered a postglacial wasteland. Getting there first, rather than in large numbers, is the key to the dominance of these founders in our British genes, and this applies to both women (mitochondrial DNA) and men (Y chromosome). Gordon Brown's Britishness needs to be extended back by at least 13,000 years from the familiar world of Celts and Romans to consider those who contributed most to our common heritage. A further 700,000 years needs to be added if we are to understand the full evolutionary picture.

And what is the policy implication? This concerns an overhaul of the UK national curriculum, where presently the debate on British identity starts with the Middle Ages. As a result, education is being denied access to scientific progress.

These three books show why we can no longer ignore our earliest ancestry in deciding who the British think they are. It is time to celebrate those first economic migrants, because that is who we are. ■

Clive Gamble is in the Department of Geography, Royal Holloway, Egham TW20 0EX, UK. His latest book, *Origins and Revolutions: Human Identity in Earliest Prehistory*, will be published in 2007.

Treasure islands

H. & J. ERIKSEN

The Socotra islands, in the Arabian Sea off the Horn of Africa, are home to many plant and animal species found nowhere else, including the Socotra sunbird shown here. First settled by man several thousand years ago, this unique environment is populated by a small group of fishers and pastoralists. In 2003, Socotra became a UNESCO Man and Biosphere Reserve. *Socotra: A Natural History of the Islands and their People* by Catherine Cheung and Lyndon DeVantier (Socotra Conservation Fund/Odyssey, £39.50) is the first full natural history of the flora, fauna and people of these islands.

A physics travelogue

From Clockwork to Crapshoot: A History of Physics
by Roger G. Newton
Belknap Press: 2007. 352 pp. $29.95, £19.95, €27.70

David Lindley

If you have just an afternoon to spare for your first visit to the British Museum in London, you have a choice to make. You can trot smartly up and down the corridors, trying to glimpse as many items as possible, or you can choose to linger thoughtfully in a handful of rooms, hoping to absorb a sense of the entire collection's scope. In his role as tour guide to the complete history of physics, Roger Newton seems to have had trouble deciding which strategy to adopt. Sometimes he pauses to reflect on the meaning and significance of the most crucial exhibits; at other times he seems determined to march briskly down the centuries, ticking off names and discoveries great and small with bewildering haste. As a result, the truly interesting perspectives that he points out along the way get lost in the confusion.

From Clockwork to Crapshoot begins by defending Aristotle against the bad press he sometimes gets in histories of science. While Plato mused abstractly about the ideal nature of things, Aristotle turned his attention to the 'efficient causes' of empirical phenomena — meaning, in a nutshell, that if something happens, there must be something else that makes it happen. That is a modern philosophy of science, but in his specifics, Aristotle was mostly wrong. It was medieval scholars, rediscovering Aristotle from Arab writers, who treated his writings as a revealed truth, insisting on scrupulous adherence to his incorrect explanations but failing to grasp his style of reasoning.

After dropping in on Roger Bacon, William of Ockham and Nicole Oresme, we're onto Copernicus, Galileo, Kepler and Newton — the beginning of science as we now understand the term. This is familiar territory, and although the author's travelogue is fluent and intelligent, the narrative interest starts to flag. With the basic method of science settled, the story is one of advancing enlightenment on many fronts, and our guide is determined to give at least a brief wave to everyone who contributed. Taken individually, his sketches of Laplace, d'Alembert and Gauss, of Henry, Faraday and Maxwell, and of Rumford, Joule and Clausius, are engaging enough. Thrown at the reader one after the other, they become rather wearisome.

The story picks up again when the author tackles the emergence of statistical mechanics and then quantum mechanics. As the book's title suggests, the evolution away from strict determinism into a world governed by laws of probability marked a tectonic shift in the foundations of science. Quantum theory raised questions about the meaning of physical reality that remain unresolved today. And of late, Roger Newton suggests, Plato is staging a comeback against Aristotle. Now that we have a pretty good understanding of how electrons and other particles behave, we are returning, in attempts to find a 'theory of everything', to the deeper problem of understanding why

these particles exist, and what determines their qualities.

The author is most compelling when he tackles these broad historical trends in the scope and purpose of physical theorizing. But these large themes only occasionally come to the fore. It is also unclear what kind of reader he imagines he's writing for. Discussing the emergence of quantum mechanics, for example, he observes in passing that Dirac's formulation derived more from the poissonian than the hamiltonian version of classical mechanics, a remark that will mean something only to those who already know what it means.

Readers with some general knowledge of the development of physics will find in Roger Newton a companionable guide who points out familiar and vaguely remembered landmarks and offers occasional illuminating commentary. If his aim was to enlighten a less well-versed audience, he could have said more by saying less. ■

David Lindley is a freelance writer in Alexandria, Virginia, USA.

On the right path

The Best of All Possible Worlds: Mathematics and Destiny
by Ivar Ekeland
Chicago University Press: 2006. 191 pp. $25

Joseph Mazur

Thomas Aquinas argued that evil helps the good in the world. St Augustine maintained that God brought evil into the Universe to bring about a greater good. And Gottfried Leibniz, after confessing in his *Theodicy* that God was free to create a world without evil, asserted that the best plan for a Universe is "not always that which seeks to avoid evil, since it may happen that the evil is accompanied by a greater good". He concluded that of all the worlds God could have created, the one we live in is "the best of all possible". We remember this phrase best through *Candide*, Voltaire's satire of Leibniz's philosophy: "If this is the best of possible worlds," asks Dr Pangloss, Candide's teacher, "what then are the others?"

How much evil is needed to maximize good? It seems that God has chosen one world from an infinite collection of possibilities by seeking to minimize evil under the constraint of maximizing good. The eighteenth-century French philosopher Pierre-Louis Moreau de Maupertuis gave us the principle of least action: in all natural phenomena, a quantity called 'action' — for him, the product of mass, distance travelled and velocity — tends to be minimized. In his view, God, being the supreme mathematician, had created the "best of all possible worlds" by insisting that everything in it obey the principle of least action, an economy of effort — a metaphysical rule designed to support the laws of mechanics.

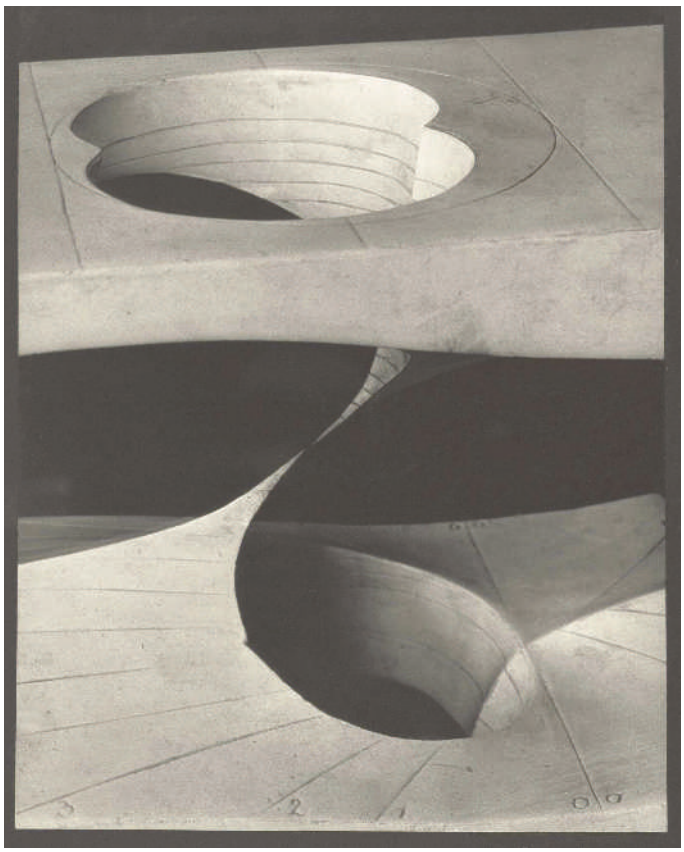
In *The Best of All Possible Worlds*, Ivar Ekeland skilfully traces the historical developments of de Maupertuis' principle as it matured from a metaphysical directive in physical two- or three-dimensional space to a mathematical principle in a conceptual space where the action is not just minimized but stopped altogether. He then tracks it further to our modern notions of randomness measured by probabilities. This complex story can be read with a minimum of effort, and we are left feeling that Maupertuis' principle works, even though we know that randomness is hardly compatible with minimizing actions. Ekeland — a distinguished mathematician and director of the Pacific Institute for Mathematical Sciences in Vancouver — goes on to say: "If there is a God, he has left no tracks in the laws of physics; or if he has, he has covered them up very well."

The real question behind Ekeland's magnificent book is this: how does nature do it? Of all the possible actions, from travelling photons to rolling billiard balls, how does nature choose what path to follow? You are reading a book review of Ivar Ekeland's *The Best of All Possible Worlds*, but how did you get to be reading it at this precise moment and in this place? Was it a dictate of nature that led to this action? Was it the perturbations caused by the proverbial butterfly flapping its wings over the Pacific? Did nature have some purpose in making you read it?

Ekeland says 'no' to all three questions. Yes, some mechanical systems are modelled by criteria of optimizing performance that could be interpreted as minimizing some kind of action or energy. But most mechanical systems of the world are unpredictable. "Since classical mechanics has dealt exclusively with integrable systems for so many years, we have been left with wrong ideas about causality," writes Ekeland; we have ignored non-integrable systems — those that do not admit exact solutions to differential equations. World events are not linear. History does not follows parallel causal chains; each event "is like the trunk of a tree, plunging a network of roots deep into the past, and raising a crown of branches high into the future". Ekeland refutes Blaise Pascal's remark that if Cleopatra had a shorter nose, the world would be very different.

Ekeland competently weaves the philosophical views of scientists through the warp of metaphysics dealing with nature's directives. We hear how Ernst Mach believed that the role of science is to explain the facts as accurately and simply as possible, and how Henri Poincaré believed that science is not really about objective reality or truth, but rather the ease and expediency of human comprehension. Over the weave, Ekeland embroiders some lively anecdotes involving illustrious individuals and great historical moments, ranging from the Peloponnesian Wars and Venetian concessions to the Hapsburg emperor Maximilian to Darwin's voyage to the Galapagos. His explanations are clear and elegant, in the brilliant, effortless manner of Richard Feynman, and his prose is fluid, exhilarating and suspenseful. I tried to put this superb book down after chapter 4 but couldn't. It was as if some compelling force of nature had a purpose, an opposing directive in the best of all possible worlds. ■

Joseph Mazur is the author of *Euclid in the Rainforest*. He teaches mathematics at Marlboro College, Marlboro, Vermont 05344 , USA.



J. PAUL GETTY TRUST/MAN RAY TRUST ARS-ADAGP

CONSTRUCTIVE MEMORY

The ghosts of past and future

A memory that works by piecing together bits of the past may be better suited to simulating future events than one that is a store of perfect records.

Daniel L. Schacter and Donna Rose Addis

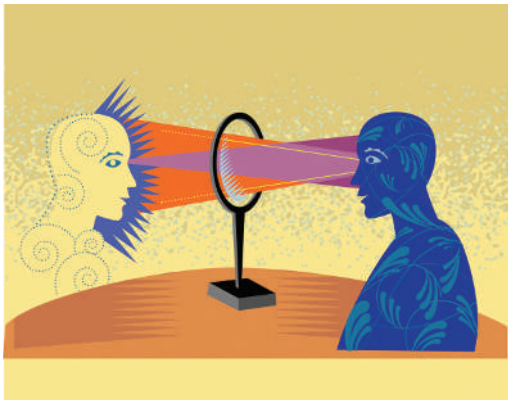
In the days after the 1995 Oklahoma City bombing, the authorities began hunting for a suspect they named John Doe 2. A mechanic had vividly recalled seeing this man with the bomber, Timothy McVeigh, at a body shop where they rented the van used to carry out the crime. But John Doe 2 was never found. Further investigations revealed that the mechanic had mistakenly recalled that an innocent man he saw the next day at the body shop — with someone who looked like McVeigh — had accompanied McVeigh the day before. The mechanic had combined accurate bits of memory from two separate episodes into a single, inaccurate recollection.

As this example highlights, memory errors can create confusion in everyday life. But for scientists studying memory, such mistakes are important because they provide critical evidence for the idea that episodic memory — the kind of memory that allows us to remember our personal experiences — is not a literal reproduction of the past, but is instead constructed by pulling together pieces of information from different sources.

Such a characterization of memory is hardly new: the psychologist Frederic Bartlett argued for it in his famous 1932 book *Remembering*. But we still understand little about how constructive remembering is achieved by the brain, and even less about the function served by such a system. Why are episodic memories created by piecing together bits of information, and not by simply reproducing the past as a video-recorder does? Recent research provides clues to the puzzle, and suggests a novel perspective.

One clue comes from studies indicating that memory errors can reveal the operation of adaptive rather than defective processes. For example, consider the following words: tired, bed, awake, rest, dream, night, blanket, doze, slumber, snore, pillow, peace, yawn and drowsy. When asked whether 'blanket' was on the list (a few minutes after seeing the words), most people correctly recognize that it was; when asked about 'point', they correctly remember that it was not. When asked about 'sleep', most people confidently remember having seen it — but they are wrong. They falsely recognize 'sleep' because they remember that many associated words were present, and mistakenly rely on their accurate memory for the general theme of the list.

Amnesic patients with damage to brain structures, such as the hippocampus, that are required for accurate episodic memory make fewer false-recognition errors than do control subjects, reflecting their poor memory for the theme or gist of the list. Neuroimaging studies using this word-recall procedure with normal adults have found that both true and false recognition show similar levels of brain activity in the hippocampus and other regions in the parietal and frontal lobes that are usually associated with accurate remembering.



R. TOTSUMOTO/IMAGES.COM/CORBIS

Taken together, neurological and neuroimaging studies suggest that false-recognition errors reflect the healthy operation of adaptive, constructive processes supporting the ability to remember what actually happened in the past. Many researchers believe that remembering the gist of what happened is an economical way of storing the most important aspects of our experiences without cluttering memory with trivial details. We agree. But we also see another important function for constructive memory, one that emerges from an idea that a growing number of researchers are embracing — that memory is important for the future as well as the past.

As Yadin Dudai and Mary Carruthers have discussed (*Nature* **434,** 567; 2005), people draw on past experiences in order to imagine and simulate episodes that might occur in their personal futures. When we imagine different versions of tomorrow's big meeting or what might happen during next week's trip, for example, we project ourselves into the future based on what we remember from the past. Indeed, information about the past is useful only to the extent that it allows us to anticipate what may happen in the future.

But future events are not exact replicas of past events, and a memory system that simply stored rote records would not be well-suited to simulating future events. A system built according to constructive principles may be a better tool for the job: it can draw on the elements and gist of the past, and extract, recombine and reassemble them into imaginary events that never occurred in that exact form. Such a system will occasionally produce memory errors, but it also provides considerable flexibility.

Although this hypothesis has yet to be subjected to direct experimental tests, there is supporting evidence. Some amnesic patients who remember little from their personal past have similar problems envisaging their future. For example, one patient who suffered from total loss of episodic memory as a result of a head injury was also unable to envisage events in his personal future, including 'this afternoon', 'tomorrow' or 'next summer'. Yet this patient performed well on tests requiring him to imagine non-personal information, such as the shapes of animals or the relative sizes of objects.

Other studies indicate that severely depressed patients think about both past and future events in an overly general, non-specific manner. Neuroimaging studies from our laboratory and others reveal striking commonalities in the brain networks that are activated when people remember past episodes and imagine future ones — for example, the hippocampus may recombine details from past events into novel future events.

For more than 100 years, memory has been the object of experimental studies that have focused almost exclusively on its role in preserving and recovering the past. We think it is time to try to understand some of memory's errors by looking to the future. ■

Daniel L. Schacter and Donna Rose Addis are in the Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, Massachusetts 02138, USA.

FURTHER READING
Schacter, D. L. *The Seven Sins of Memory* (Houghton Mifflin, 2001).
Suddendorf, T. & Corballis, M. C. *Genet. Soc. Gen. Psychol. Monographs* **123,** 133–167 (1997).
Tulving, E. *Annu. Rev. Psychol.* **53,** 1–25 (2002).

CONCEPTS

NEWS & VIEWS

PLANETARY SCIENCE

Titan's lost seas found

Christophe Sotin

When the Cassini spacecraft found no methane ocean swathing Saturn's moon Titan, it was a blow to proponents of an Earth-like world. The discovery of northern lakes on Titan gives them reason for cheer.

The saturnian moon Titan is the second largest satellite in the Solar System, trumped only by Jupiter's Ganymede. It is the only Solar System satellite with a dense atmosphere, which produces a surface pressure 1.5 times that at Earth's surface. And it shares with Earth the peculiarity that nitrogen is the principal component of its atmosphere. The list of similarities does not end there, and, as Stofan *et al.* report[1], it has just been augmented. The authors' account of what seem to be lakes at high northern latitudes on Titan appears on page 61 of this issue.

The lakes are not formed of water, as they would be in earthly climes, but of the second most abundant component of Titan's atmosphere, methane (CH_4). The bounteous presence of methane and aerosols in Titan's enveloping cloak hides the surface of the moon at visible wavelengths. For this reason, little was known about Titan's inner life before the arrival of the joint NASA/European Space Agency Cassini–Huygens mission in the Saturn system on 1 July 2004.

The lifetime of methane is short on geological timescales: the molecule lasts some tens of millions of years before it becomes dissociated by sunlight. Before the first results arrived from Cassini–Huygens, two hypotheses had been advanced to explain how, in the face of this slow depletion, Titan replenishes its atmospheric methane. First, that a methane-rich hydrocarbon ocean covers Titan's solid surface, and supplies the atmosphere in a cycle of evaporation and condensation[2]. Alternatively, that underground methane reservoirs exist just below the surface or deep in Titan's interior, which deliver methane to the outside through 'cryovolcanic' processes or when the surface is punctured by meteorite impacts. The first of these pictures was the more popular, and would have made Titan even more similar to Earth, with the extraordinary shared feature of a surface ocean. The Huygens probe, which was to be released by the Cassini spacecraft as it flew past Titan, was designed to survive for several minutes on reaching the assumed ocean's surface.

On 26 October 2004, a couple of months before it did release Huygens, Cassini performed its first close fly-by of Titan, skimming

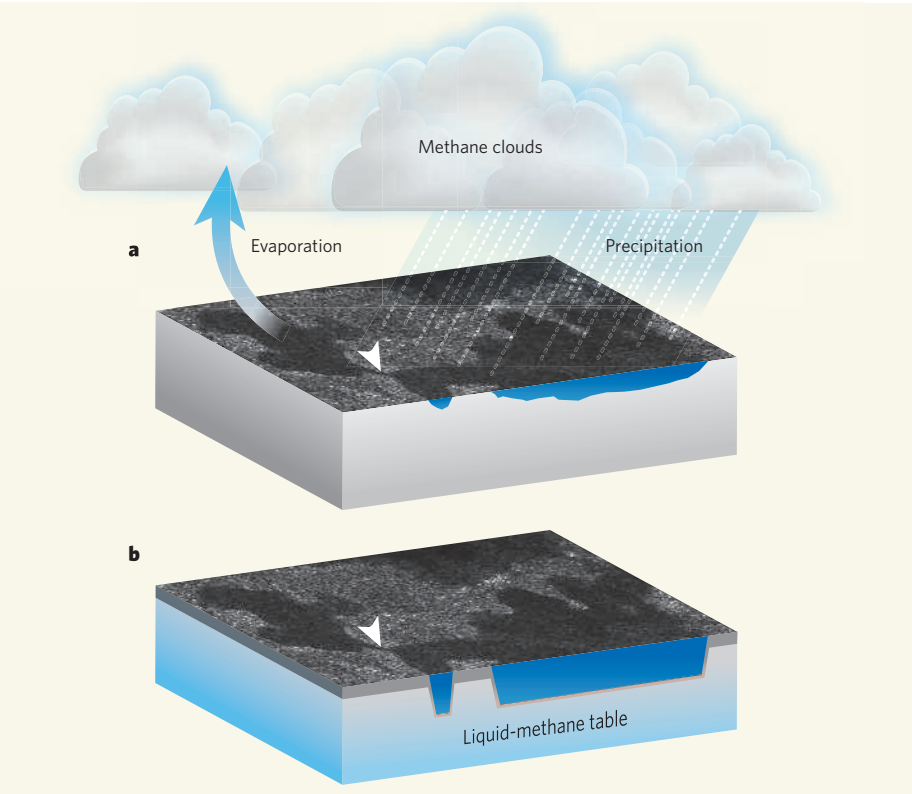


Figure 1 | Routes to Titan's lakes. The lakes discovered by Stofan *et al.*[1] might be either **a**, filled by methane rain, either directly or through river inflow, or **b**, in depressions filled from an underground liquid-methane table. The surface images are taken by the Cassini spacecraft, and seem to show liquid bodies, two of which are connected by a channel (arrow). (Cassini image taken from ref. 1.)

its atmosphere 1,174 kilometres from the surface. Three remote-sensing instruments trained on the surface failed to detect a global ocean. What they detected instead was even more fascinating: impact craters, mountains, cryovolcanoes, dunes and river beds[3]. The lack of a global ocean and the discovery of these surface features, together with characteristics of Titan's atmosphere such as its nitrogen and carbon isotopic ratios[4], strongly implied that the source of the atmospheric methane was internal. With Stofan and colleagues' discovery[1] of lakes at northern latitudes, the pendulum starts to swing the other way once more.

Their report is based on observations made by Cassini's radar instrumentation in July 2006. These revealed around 75 radar-dark patches, ranging from 3 to 70 km in size, at latitudes between 70° N and 83° N. Such dark areas are characteristic of very smooth surfaces. Their liquid nature is inferred from the presence of channels leading to them, seeming to indicate that rivers supply at least part of the liquid. Although the composition of the liquid cannot be determined from radar observations, methane is the most plausible candidate: it is one of few molecules to be liquid under the conditions of Titan's surface.

The findings provide further strong evidence, complementary to that inferred from the river beds observed by the Huygens probe during its descent[5], that methane on Titan plays the role of water on Earth: liquid methane evaporates; the vapour eventually condenses; and rainfall

replenishes the surface liquid (Fig. 1a). An alternative is that the surface liquid comes from a 'liquid-methane' table that fills in the topographic lows of the surface (Fig. 1b). By comparison with the morphologies of terrestrial lakes, the authors suggest that the depressions could be impact craters, volcanic calderas or the sinkholes (dolines) characteristic of karst landscapes. Such landscapes are formed on Earth by the dissolution of carbonate rocks by rainwater.

The fact that lakes are found only at high latitude in Titan's northern hemisphere seems to indicate that they expand during the winter and shrink in the summer as a result of increased evaporation (it is winter in Titan's northern hemisphere at the moment). This cycle is linked to the 29.5 years it takes Saturn to orbit the Sun. On longer timescales, Titan's atmosphere might also be replenished in methane by cryovolcanic activity, as geomorphological features observed by Cassini imply[6].

The Cassini mission is now halfway to the end of its nominal mission, and the detailed morphology of Titan's surface is becoming steadily clearer at each fly-by. Like a giant puzzle, our understanding of Titan's dynamics is coming together as we connect the pieces. There will undoubtedly be other discoveries during the next 22 Titan fly-bys, the next of them due on 13 January. By the end of the planned mission, however, Cassini's radar will have covered only 15% of Titan's surface, and its Visual and Infrared Mapping Spectrometer just a few per cent, at a resolution of less than a kilometre per pixel. An extended mission, currently under discussion, is necessary to gain better coverage of Titan's surface. Cassini's optical and infrared instrumentation could then also be used to monitor the evolution of the northern lakes — currently shrouded in the darkness of the titanian winter — as they enter the Saturn system's summer season next year.

Stofan and colleagues' findings[1] add to the weight of evidence that Titan is a complex world in which the interaction between inner and outer layers is controlled by processes similar to those that must have dominated the evolution of any Earth-like planet. Indeed, as far as we know, there is only one planetary body that displays more dynamism than Titan. Its name is Earth. ■

Christophe Sotin is at the Laboratoire de Planétologie et Géodynamique, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes cedex 3, France, and a visiting scientist at the Jet Propulsion Laboratory , Pasadena, California, USA.
e-mail: christophe.sotin@univ-nantes.fr

1. Stofan, E. R. *et al. Nature* **445,** 61–64 (2007).
2. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. *Science* **222,** 1229–1230 (1983).
3. Elachi, C. *et al. Science* **308,** 970–974 (2005).
4. Niemann, H. *et al. Nature* **438,** 779–784 (2005).
5. Tomasko, M. G. *et al. Nature* **438,** 765–778 (2005).
6. Sotin, C. *et al. Nature* **435,** 786–789 (2005).

NEUROBIOLOGY

Scent secrets of insects

Rachel I. Wilson

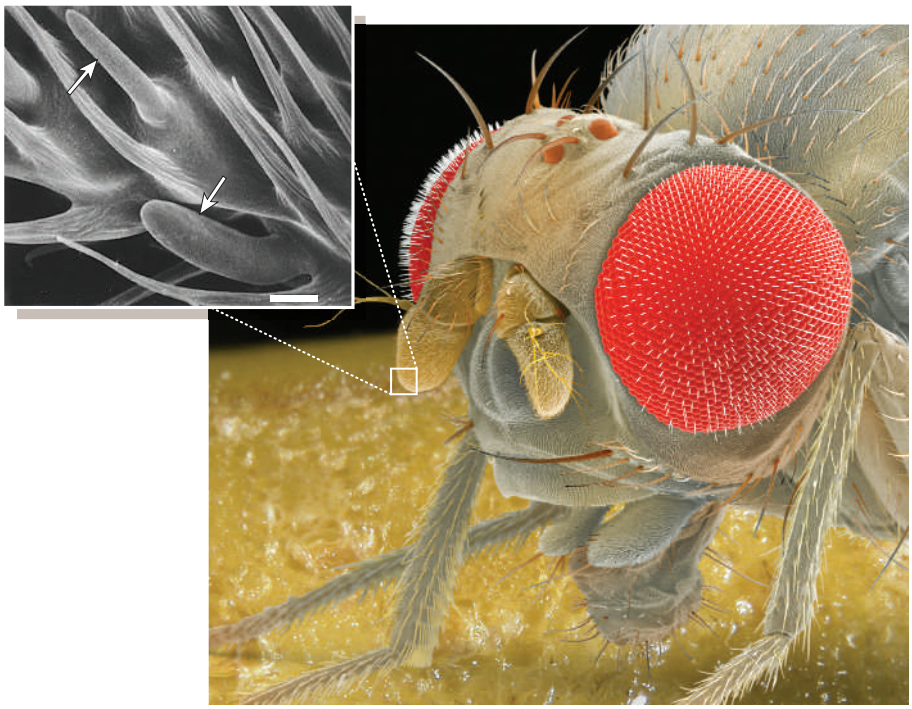
The perception of carbon dioxide provides insects with sensory data on their environment, and informs many insect behaviours. It seems that this sense relies on two dedicated neural receptors.

We inhabit a different sensory universe from that of many of the animals around us. We are deaf to high-pitched sounds that dogs perceive, blind to ultraviolet light that honeybees see, and numb to electric fields that sharks feel. And there is a world of chemicals swirling around us that we cannot smell, but that carry pungent signals for other species.

One such signal is carbon dioxide, which many insects sense through specialized neurons. At very high concentrations, CO_2 can be perceived by the human nose — recall the last time you opened a can of carbonated soda and sniffed before sipping. But many insects are exquisitely sensitive to concentrations we never notice, and monitoring CO_2 levels in the environment is crucial to many insect lifestyles. For example, some ticks can detect CO_2 fluctuations as small as 20 parts per million[1]; for a blood-sucking insect, elevated CO_2 means that a potential host animal might be nearby. Inside a beehive, high CO_2 levels mean that ventilation is needed to improve air quality[2]. But the molecular basis of CO_2 sensing in insects has remained a mystery.

On page 86 of this issue, Jones *et al.*[3] report the identification of two receptor proteins that together are required for CO_2 perception in the fruitfly *Drosophila melanogaster*. This advance contributes to our understanding of the way in which very small volatile molecules such as CO_2 are sensed by cells, and it has the potential to facilitate innovative insect control strategies.

The perception of volatile chemicals begins when molecules in the air interact with receptor proteins on the surface of olfactory neurons. In fruitflies, these neurons reside in two specialized organs, the antennae and the maxillary palps (Fig. 1). Among the approximately 1,200 olfactory neurons in each antenna are 45 CO_2-sensitive neurons. Whereas most other antennal neurons are activated by several different volatile chemicals, CO_2-sensing neurons in fruitflies respond to just this one stimulus[4]. These neurons do not express any of the olfactory receptor genes that are responsible for sensing other odours. Instead, previous work showed that they express a receptor that is similar to taste receptors, so it was classed as a gustatory receptor (GR)[5,6], even though it is evidently unrelated to taste. The gene encoding



J. RIESGO-ESCOVER & J. CARLSON

A. SYRED/SPL

Figure 1 | Sensing carbon dioxide. The fruitfly *Drosophila* has carbon dioxide-sensitive neurons on its antennae (inset, arrows). Inset: scale bar, 0.2 μm; reproduced from ref. 12.

this receptor, *Gr21a*, was shown to be expressed only in CO_2-sensitive neurons[7,8]. This suggested that *Gr21a* might be involved in CO_2 sensation. However, *Gr21a* alone was insufficient to confer sensitivity to CO_2 when it was expressed in other neurons, implying that an essential factor in the process was still missing.

Jones *et al.*[3] reasoned that the missing partner might also be similar to gustatory receptors. They discovered that one such gene, *Gr63a*, is indeed expressed with *Gr21a* in CO_2-sensitive neurons. Moreover, when these two genes were expressed together in another antennal neuron (a conventional olfactory neuron), they conferred robust responses to CO_2 on that cell. However, neither gene alone was sufficient to produce CO_2 sensitivity.

Next, the authors genetically engineered flies that lacked the *Gr63a* gene. In these 'knockout' flies, the neurons that normally respond to CO_2 were completely unresponsive. And whereas fruitflies normally avoid CO_2, the knockout flies were indifferent to this odour. This state of affairs was reversed by adding back a *Gr63a* gene to the knockout flies, demonstrating that loss of this gene was indeed responsible for the sensory deficit.

Together, these results demonstrate that both *Gr21a* and *Gr63a* are required for CO_2 perception in *Drosophila*. The simplest scenario is that the two receptors form a complex that binds to CO_2. It is possible, however, that other molecules are also required. If so, these components must be present in conventional olfactory neurons, because *Gr21a* and *Gr63a* were together sufficient to confer CO_2 sensitivity when expressed in an arbitrary olfactory neuron elsewhere in the antenna.

Another open question is whether this putative receptor complex actually binds to CO_2. In vertebrates, elevation of CO_2 excites neurons that modulate breathing rhythms, increasing respiration and helping to clear CO_2 from the blood. This response is not, however, mediated by a direct action of CO_2. Instead, the neurons involved are activated by changes in pH that are secondary to CO_2 elevation[9]. A similar process might be occurring in the *Drosophila* antenna. If the receptor complex does bind to CO_2 directly, it will be interesting to discover what this binding site looks like. Many cellular responses to gases are mediated by metalloproteins, suggesting that a metal cofactor might have a role in this complex.

Understanding how this receptor complex interacts with CO_2 should also shed light on the unusual response properties of CO_2-sensitive neurons in insects[10,11]. Compared with conventional olfactory neurons, these neurons are unusually insensitive to the velocity of air flow around the antenna. They signal concentration steps independently of background CO_2 levels, and respond to CO_2 increases and decreases in a remarkably symmetric way. Their concentration–response function is also nearly linear at concentrations near the typical ambient level of CO_2. Considered as tiny chemical sensors, these neurons are wonders of natural engineering.

Finally, the discoveries reported by Jones *et al.*[3] have the potential to contribute to disease prevention. The most dangerous animals on Earth are in fact mosquitoes — mosquito-borne diseases cause more than a million deaths annually around the world. And like other blood-sucking insects, mosquitoes use CO_2 to locate their hosts. Jones *et al.* show that the mosquito relatives of *Gr21a* and *Gr63a* are co-expressed in the mosquito maxillary palp, a structure known to be the locus of CO_2 sensation in these insects. If this molecular insight permits the design of novel mosquito deterrents, it could have a major impact on global health. ■

Rachel I. Wilson is in the Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mail: rachel_wilson@hms.harvard.edu

1. Stange, G. & Stowe, S. *Microsc. Res. Tech.* **47,** 416–427 (1999).
2. Nicolas, G. & Sillans, D. *Annu. Rev. Entomol.* **34,** 97–116 (1989).
3. Jones, W. D., Cayirlioglou, P., Kadow, I. G. & Vosshall, L. B. *Nature* **445,** 86–90 (2007).
4. de Bruyne, M., Foster, K. & Carlson, J. R. *Neuron* **30,** 537–552 (2001).
5. Clyne, P. J., Warr, C. G. & Carlson, J. R. *Science* **287,** 1830–1834 (2000).
6. Scott, K. *et al. Cell* **104,** 661–673 (2001).
7. Suh, G. S. *et al. Nature* **431,** 854–859 (2004).
8. Couto, A., Alenius, M. & Dickson, B. J. *Curr. Biol.* **15,** 1535–1547 (2005).
9. Feldman, J. L., Mitchell, G. S. & Nattie, E. E. *Annu. Rev. Neurosci.* **26,** 239–266 (2003).
10. Stange, G. *J. Comp. Physiol. A* **171,** 317–324 (1992).
11. Grant, A. J., Wigton, B. E., Aghajanian, J. G. & O'Connell, R. J. *J. Comp. Physiol. A* **177,** 389–396 (1995).
12. Riesgo-Escovar, J. R., Piekos, W. B. & Carlson, J. R. *J. Comp. Physiol. A* **180,** 151–160 (1997).

BIOORGANIC CHEMISTRY

A sweet synthesis

Linda C. Hsieh-Wilson

Peptides and proteins with sugars attached have many desirable biological properties, but their chemical synthesis is a technical challenge. An ingenious take on an old idea might simplify things considerably.

Part of what distinguishes us from bacteria is that the proteins in our bodies are decorated with elaborate arrays of sugars. Protein glycosylation — the attachment of sugars to the amino-acid building-blocks of proteins — plays a crucial role in such diverse processes as protein folding, cell–cell communication and viral invasion of cells. Yet it is conspicuously absent in many simple, unicellular organisms. Understanding the roles of these sugars and how their complex, disparate structures modulate the activities of proteins has been a long-standing challenge. Reporting in the *Journal of the American Chemical Society*[1], Brik and colleagues bring us a step closer to this goal by devising a clever strategy for generating glycopeptides — short sequences of amino acids with sugars attached — that may one day permit the tailored synthesis of glycoproteins.

Glycopeptides and glycoproteins are notoriously difficult to obtain as pure compounds, because they are naturally expressed as inseparable mixtures of different structures (glycoforms) that bear various sugars. This complexity makes it difficult to study how any specific glycoform affects a protein's function, which in turn complicates efforts to generate protein-based medicines. Indeed, most therapeutic glycoproteins are sold as mixtures of glycoforms, the active components of which are often unknown. One approach to solving this problem is to use chemical synthesis to create single structures.

Brik and colleagues[1] have now developed a strategy for assembling glycopeptides using a process known as peptide ligation. In their method, one peptide is attached to another that incorporates a modified sugar. A unique feature of this approach is that the sugar assists the process by positioning the two peptides in close proximity to each other. Traditional glycopeptide synthesis is cumbersome, requiring excesses of reagents to drive reactions to completion, and often producing low yields of the desired products. Furthermore, strategies involving 'protecting groups' have been necessary to mask reactive chemical groups that do not participate directly in the reaction sequence. These requirements increase the complexity and the cost of glycopeptide synthesis. But by actively engaging a sugar in the ligation process, Brik *et al.* demonstrate that a variety of glycopeptides can be made in just a few steps and in high yield, without the need for protecting groups.

The authors' strategy is a clever twist on a well-established method for peptide synthesis known as native chemical ligation[2]. In this process, two peptide fragments are joined together to form a larger fragment via a two-step mechanism. The first step involves the transient formation of a thioester bond between the two fragments (Fig. 1a), mediated by a reactive sulphur atom on one of the fragments. The resulting intermediate then undergoes a rapid, spontaneous rearrangement to form a peptide bond. The net result is the direct connection of two peptide fragments to form a

50 & 100 YEARS AGO

50 YEARS AGO

Among the numerous well-known scientists who were born in 1857...Ronald Ross is widely known, for the story of his long and patient attempts to identify the carrier of malarial fever has often been written... An outstanding centenary of the present year is that of the birth of Heinrich Rudolf Hertz, the German physicist, who was the first to detect electromagnetic waves in free space and measure their velocity... Elwood Haynes (1857–1925) is another American who should be remembered this year. He discovered several important alloys, including tungsten chrome steel, and in 1919 filed a patent for stainless steel... The question of the inheritance of acquired characteristics has recently received much attention. An early worker in this field of research was the Danish botanist W. L. Johannsen (1857–1927). One of the founders of modern research in heredity, he introduced the terms 'pure line', as well as 'gene', 'genotype' and 'phenotype'.
From *Nature* 5 January 1957.

100 YEARS AGO

In a recent note attention was directed to the recent renewal of experiments with Count Zeppelin's latest airship on the Lake of Constance... The 1906 Zeppelin airship...is 11 metres high, and each of the two cars can hold four persons, besides having a separate motor. The author states that with both motors working simultaneously a speed of 15 metres per second, or 54 kilometres per hour, can be maintained for sixty hours with the quantity of benzene the machine will carry... The advantages of the Zeppelin airship are more or less counterbalanced by the present necessity of using a sheet of water for starting and landing. Apart from the uses of such a machine in warfare, its applications in time of peace to the meteorological survey of the atmosphere are contemplated.
From *Nature* 3 January 1907.

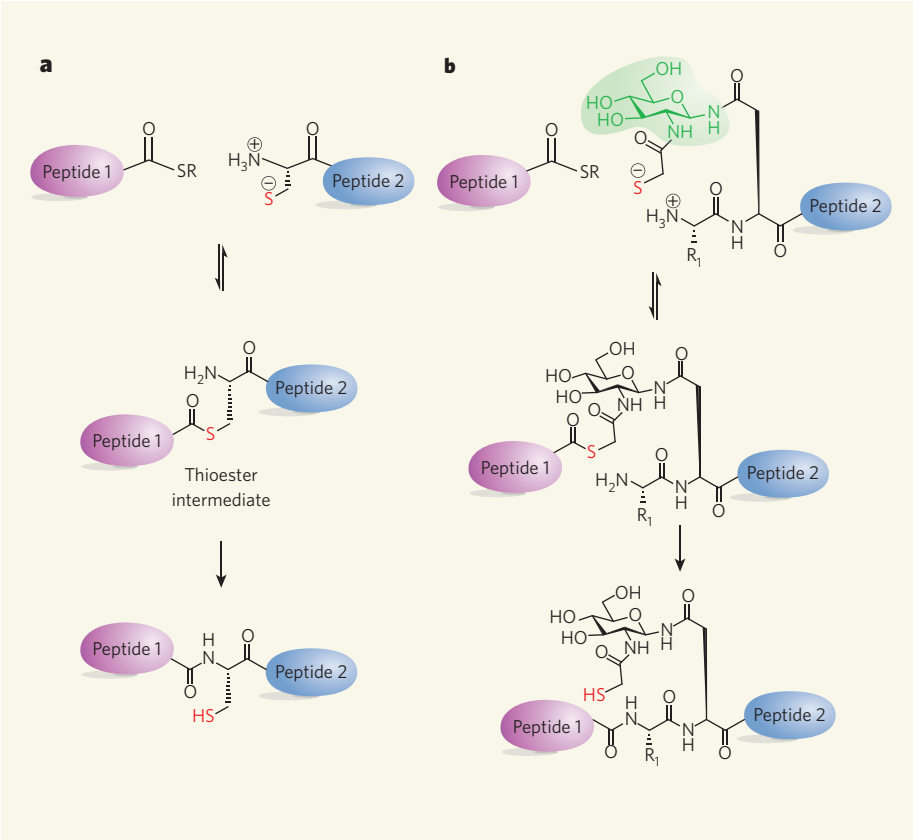


Figure 1 | Glycopeptide synthesis. a, Native chemical ligation is a well-established method for preparing peptides. A reactive sulphur atom (red) on the side-chain of a cysteine amino acid attacks another peptide (where R is typically a phenyl ring), producing a thioester intermediate that spontaneously rearranges to yield a peptide bond. **b**, Brik *et al.*[1] have modified this method to prepare glycopeptides, in which sugars are attached to peptide chains. A reactive sulphur atom (red) attached to an appended sugar (green) acts as a surrogate for the cysteine side-chain. Peptide bonds can thus be formed between a greater variety of amino acids. R_1 represents an amino-acid side-chain.

larger polypeptide. Moderately sized proteins have been produced in this way by sequential ligation of several peptide fragments, or through the coupling of a peptide to a larger protein fragment. Crucially, native chemical ligation provides exquisite control over the protein structure being formed, and allows the incorporation of various useful groups — such as synthetic amino acids, biophysical probes or stable isotopes of atoms used for structural studies — into selected sites within proteins[3,4].

Building on this approach, Brik and colleagues[1] attached a reactive sulphur group to a sugar within a peptide (Fig. 1b). In a process similar to the two-step mechanism for native chemical ligation, the authors reacted this sulphur group with a second peptide to form a thioester intermediate. This intermediate subsequently rearranges to give the desired product, in which the two starting materials are linked by a peptide bond. This strategy[1] has several remarkable features. Native chemical ligation requires cysteine — a sulphur-containing amino acid — to be at the reacting end of one of the peptides being joined together. By placing a reactive sulphur group on the sugar of a glycopeptide, rather than in an amino acid, the authors circumvent this requirement, thus allowing bonds to be formed between a broader range of amino acids.

Moreover, a surprisingly wide array of amino acids is tolerated at the reaction site, thus permitting access to glycopeptides that are difficult to synthesize using other methods. Amino acids with small side-chains and those (such as histidine or aspartate) with side-chains that can serve as a base in the ligation pathway are favoured substrates in the reaction. Finally, the sulphur atom on the sugar provides a convenient handle for subsequent chemical manipulation — for example, it can be removed to give a naturally occurring sugar, reacted to append fluorescent dyes or other groups to the glycopeptide, or elaborated to form more complex sugars by using glycosyltransferase enzymes[1].

Further investigations are needed to assess the full scope of Brik and colleagues' reaction[1] and its potential application to glycoprotein synthesis. Nonetheless, the emergence of this and other methods[5–8] for constructing pure peptides and proteins with sugars installed at preselected sites has many implications. For example, such techniques could transform the way therapeutic glycoproteins are discovered, developed and manufactured. Many of these proteins are obtained only as a mixture of glycoforms, just a fraction of which may

be biologically active[9]. But if drug-regulation authorities start to impose stringent regulations on glycoproteins (as they currently do for traditional 'small molecule' drugs, where the purity of the active form is paramount), then single glycoforms will be required. Furthermore, the ability to fine-tune the biological properties of therapeutic proteins by modifying their attached sugars could lead to exciting advances in drug discovery.

More fundamentally, having access to pure glycoproteins would help to elucidate the role of specific sugars in regulating protein structure and function. This could help to reveal how bacteria manage without these sweet appendages. Brik and colleagues' method[1] for making pure glycopeptides (and possibly glycoproteins) is truly a milestone achievement in this rapidly developing field. ■

Linda C. Hsieh-Wilson is at the California Institute of Technology and Howard Hughes Medical Institute, Division of Chemistry and Chemical Engineering, 1200 East California Boulevard, Pasadena, California 91125, USA.
e-mail: lhw@caltech.edu

1. Brik, A. *et al. J. Am. Chem. Soc.* **128,** 15026–15033 (2006).
2. Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. H. *Science* **266,** 776–779 (1994).
3. Dawson, P. E. & Kent, S. B. H. *Annu. Rev. Biochem.* **69,** 923–960 (2000).
4. Muir, T. W. *Annu. Rev. Biochem.* **72,** 249–289 (2003).
5. Hamilton, S. R. *et al. Science* **313,** 1441–1443 (2006).
6. Warren, J. D., Miller, J. S., Keding, S. J. & Danishefsky, S. J. *J. Am. Chem. Soc.* **126,** 6576–6578 (2004).
7. Macmillan, D. & Bertozzi, C. R. *Angew. Chem. Int. Edn* **43,** 1355–1359 (2004).
8. Zhang, Z. W. *et al. Science* **303,** 371–373 (2004).
9. Haselbeck, A. *Curr. Med. Res. Opin.* **19,** 430–432 (2003).

DEVELOPMENTAL BIOLOGY

This worm is not for turning

Henry Gee

Molecular investigations of the origin of the dorso-ventral axis in an obscure marine invertebrate illuminate one of the longest-running debates in evolutionary biology — that over the origin of vertebrates.

Vertebrates are so different from other creatures that discovering their origins within the animal kingdom has always been problematic. But molecular, developmental and genomic work on the sometimes obscure invertebrate relatives of vertebrates is prompting a re-evaluation of this vexed topic.

As they recount in *PLoS Biology*, Lowe *et al.*[1] have been looking at the expression of genes associated with the specification of the dorso-ventral body axis — which surface becomes the upper (back) body surface and which the lower (belly) — in *Saccoglossus kowalevskii*, a worm-like member of the hemichordates. This is a group that is distantly related to the chordates, the larger group to which vertebrates themselves belong (Fig. 1). The authors find that the dorso-ventral axis in hemichordates is specified in a similar way to that in other animals. But this axis is decoupled from the development of the central nervous system — a later, chordate elaboration not found in hemichordates. This implies that the rules governing dorso-ventral axis formation are ancient and probably evolved with the first bilaterally symmetrical (bilaterian) multicellular animals.

The quest to understand the deployment of the dorso-ventral axis has been one of the most enduring themes in the study of vertebrate origins. It stems from the time of the wayward nineteenth-century savant Etienne Geoffroy Saint-Hilaire, who proposed that insects have the same basic body plan as vertebrates, only turned upside-down[2]. This notion joined a list of seemingly eccentric theories about vertebrate origins that has been lengthening ever since[3]. Another is the idea that vertebrates have independently invented a new kind of mouth on the opposite body surface to that in other animals. A third is that chordates and hemichordates evolved directly from ancestors akin to extinct, asymmetrical echinoderms (the group that includes modern sea-urchins and starfishes), some of which seem to have sported the characteristic gill slits seen today in chordates and, as it happens, hemichordates.

Molecular work has disposed of most of these ideas — but not without highlighting valuable grains of truth in each of them. For example, construction of molecular evolutionary trees[4] revitalized an old idea[5] that echinoderms and hemichordates are sister taxa, in which case some primitive echinoderms really did have gill slits, no longer apparent in modern forms. Likewise, the discovery[6] that insects have a genetic system of dorso-ventral specification similar to that of vertebrates — only inverted — gave Geoffroy Saint-Hilaire a new celebrity. Lowe *et al.*[1] build on this idea by showing that hemichordates exploit this same system in their development based on an axial polarity between two types of patterning molecule — BMP (bone morphogenetic protein) at one pole, and Chordin and its affiliates at the other. Baldly put, the dorso-ventral axis in all complex animals is determined largely by the antagonistic relationship of these two groups of agent.

In insects such as *Drosophila*, BMP is associated with what is conventionally regarded as the dorsal surface in the adult animal. In chordates, by contrast, BMP is a ventralizing agent. The inversion, however, is more apparent than real, having been determined after the fact by using the central nervous system — conventionally dorsal in chordates but ventral in insects — as a primary reference for telling which way is up.

Lowe and colleagues' work[1] on hemichordates adds welcome perspective. Because hemichordates have a diffuse nerve net rather than a central nervous system, this reference point disappears. Instead, we see that chordates differ from all other animals — hemichordates as well as insects — in the position of the mouth,

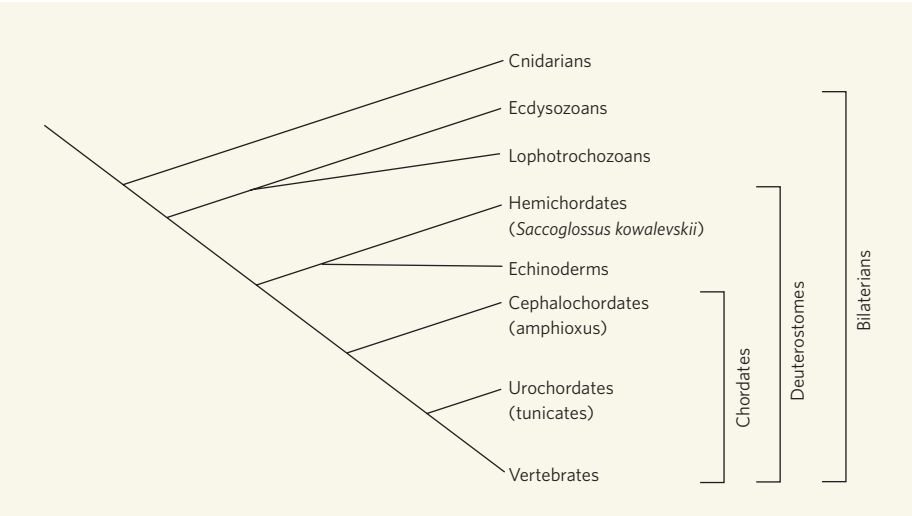


Figure 1 | Family connections. The relative position of the hemichordates in the evolutionary picture, and so of *Saccoglossus kowalevskii*, Lowe and colleagues' study subject[1]. Hemichordates, along with echinoderms (sea-urchins and allies) and chordates (which include vertebrates), are the principal members of the deuterostomes, a much larger group within the bilaterians — the bilaterally symmetrical, multicellular animals. The other principal bilaterian groups of similar rank to the deuterostomes include the ecdysozoans (insects, nematodes and others) and the lophotrochozoans (molluscs, segmented worms and others). More primitive creatures such as cnidarians (jellyfishes and others) stand outside the bilaterian grouping.

which is relocated to the Chordin side of the animal, rather than the BMP side. In this way, some of the old ideas, in which vertebrates were defined by the presence of a new mouth, had a basis in fact, however coincidentally. More seriously, this new perspective will prompt a reappraisal of the many peculiarities of the development of the mouth that are seen in lampreys (primitive, jawless vertebrates) and amphioxus (a primitive, non-vertebrate chordate). However, the central nervous systems of insects and chordates — and indeed those of all animals that have them — represent a range of solutions in which the location is governed by the BMP–Chordin axis, if not directly specified by them.

Lowe *et al.*[1] also show that in hemichordates, as in insects, dorso-ventral patterning is independent of the patterning of the anterior–posterior axis. That this is not true in chordates highlights another special feature of the latter group, perhaps associated with the development of a distinct embryonic 'organizer'. This feature has very deep roots in chordates, and is now being investigated in amphioxus[7].

The status of amphioxus itself has likewise been a matter of debate. A chordate that is conventionally regarded as the closest invertebrate relative of vertebrates, it seems[8] that it may be the most primitive known extant chordate, and thus key to our understanding of chordate innovations, including the organizer. Such a status has brought this shy creature back into a limelight it has not enjoyed since the 1930s. The amphioxus genome is not far off completion, and when it arrives it will be in select company: the genome of another primitive chordate — the tunicate *Ciona intestinalis* — has been described[9], and that of a sea-urchin, *Strongylocentrotus purpuratus*, an echinoderm long used in developmental studies, has just been announced[10].

Of course, sequencing a genome is not the same as understanding the evolution of morphological novelties. But we have come a long way since 1909, when the Linnean Society of London convened a symposium on vertebrate origins to celebrate the golden jubilee of the publication of Darwin's *Origin of Species*, and at which one participant ruefully remarked[11]: "When we return home and our friends gleefully enquire, 'What then has been decided as to the Origin of Vertebrates?', so far we seem to have no reply ready, except that the disputants agreed on one single point, namely that their opponents were all in the wrong." ■

Henry Gee is a Senior Editor of *Nature*.

1. Lowe, C. J. *et al. PLoS Biol.* **4,** 1603–1619 (2006).
2. Geoffroy Saint-Hilaire, E. *Mém. Hist. Nat.* **9,** 89–119 (1822).
3. Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, London, 1996).
4. Halanych, K. M. *et al. Mol. Phyl. Evol.* **4,** 72–76 (1995).
5. Metchnikoff, V. E. *Zool. Anz.* **4,** 139–157 (1881).
6. De Robertis, E. M. & Sasai, Y. *Nature* **380,** 37–40 (1996).
7. Holland, L. Z. *et al. Nature* doi:10.1038/nature05472 (in the press).
8. Bourlat, S. J. *et al. Nature* **444,** 85–88 (2006).
9. Dehal, P. *et al. Science* **298,** 2157–2167 (2002).
10. Sea Urchin Genome Sequencing Consortium *Science* **314,** 941–952 (2006).
11. Stebbing. T. R. R. *Proc. Linn. Soc. Lond.* **122,** 9–50 (1910).

MATERIALS SCIENCE

Alloys go with the grain

Christophe L. Martin

How do metallic alloys solidify from their original liquid state? A study of the deformation of cooling alloys confirms what had been suspected for some time: solidifying alloys bear exciting similarities to granular materials.

Metallic alloys are in constant, ubiquitous use. Generally, we prefer them in their solid state, and, in most cases, producing them requires cooling down a high-temperature liquid. This change from liquid to solid does not usually occur spontaneously at a well-defined temperature, as it does in pure metals. Instead, a continuous transition from a fully liquid to a fully solid state takes place gradually as the alloy cools.

On page 70 of this issue[1], Gourlay and Dahle provide experimental evidence that, during this process, an alloy deforms rather like a granular material. Tools developed to model such materials should therefore allow new insights into the old problem of solidification. As well as being a confident step forward into the twilight 'mushy zone' between alloy liquid and solid phases, these findings could help to elucidate the formation of defects in economically important industrial casting processes.

In the early stages of alloy solidification, small, solid grains nucleate and move freely in the liquid phase. The result is rather like a suspension, with a characteristic fluid-like behaviour. Only towards the very end of solidification — when the solid grains have been bridged together — does the material acquire mechanical coherence and behave with the extreme viscosity expected of a solid at very high temperature. Between these two extremes lies the mushy zone (Fig. 1). As this zone comprises an assembly of individual grains interacting with each other through their contacts, it is natural to conjecture that its behaviour is similar to that of a granular material. But proving this has been difficult.

Gourlay and Dahle[1] conduct their work on aluminium and magnesium alloys. Gathering experimental data on the deformation of granular materials is not easy even at room temperature. Working with materials that solidify at high temperatures (500–700 °C), and that have an extremely reactive liquid phase, is even less trivial.

The authors focus on the shear behaviour of alloys at solid fractions above 30%. Shearing occurs when two adjacent regions of a material slide past each other. In a granular material, it induces an effect called dilatancy that was first described in 1885 by Osborne Reynolds[2]. This phenomenon explains why the area around freshly laid footprints in wet sand becomes dry: deformation of the sand underfoot forces grains to rearrange, opening up spaces into which the surrounding water can flow.

As stepping on solidifying aluminium at

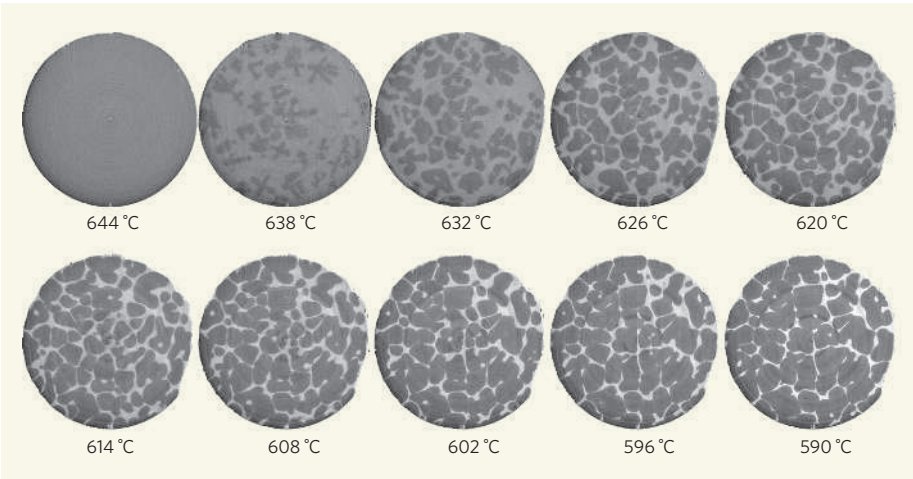


Figure 1 | Through the mushy zone. An aluminium–copper alloy gradually solidifies from the fully liquid state (above a temperature of 644 °C)[4]. The growing solid grains appear darker than the liquid phase. Gourlay and Dahle[1] show experimentally that processes of deformation in this mushy mixture are similar to those in granular materials.

600 °C is less than pleasant, Gourlay and Dahle measured shear stresses in their alloys using an instrument known as a rheometer. The authors modified their rheometer to measure the rise, as shearing proceeds, of a solid component that floats on the initially liquid alloy — equivalent to measuring the inrush of water into the shearing area under a footprint. The authors observed that, after an initial near-linear increase in resistance to shear, the solid component rises. A decrease in the stress caused by further applied shear (an effect known as strain softening) follows, with a concomitant, more gradual volume expansion. These are characteristic behaviours of granular materials.

Another trait typical of granular materials is that they develop shear bands when deformed. These are well-defined regions of intensely sheared material in which the solid grains are significantly less densely packed. A nice feature of solidifying metallic alloys, compared with more standard granular materials, is that they can be quenched — rapidly cooled — to 'freeze' the microstructure at a given point. The observation of this microstructure can reveal important information about what happened when the alloy was a mixture of solid and liquid.

The quenched microstructures observed by Gourlay and Dahle show evidence of shear bands that are typically ten grains wide. Grain size can be controlled by, for example, changing the imposed cooling rate. The authors were thus able to confirm this result for different grain sizes from 10 to 500 μm.

Shear bands that form early on in the solidification process spell bad news for the solidifying alloy, and can have significant repercussions in industrial applications. The extra liquid that drains into the band remains as solidification proceeds. This leads to segregation of the solid and liquid phases, and thus to a heterogeneous composition in the material. Worse still, later on in the solidification process this can trigger cracks — 'hot tears' — that are among the most severe defects encountered in casting and welding processes.

The scenarios sketched by Gourlay and Dahle[1] require confirmation by further experiments, as well as numerical simulations that can throw similarities with phenomena in granular materials into sharper relief. The authors' analysis is post-mortem, in the sense that it is carried out when the alloy has fully solidified. It needs verification and consolidation through experiments to track the evolution of microstructures during the actual formation of shear bands. X-ray microtomography, a technique that reveals the three-dimensional structure of materials, is an excellent candidate that has already shown the wealth of information it can provide both on standard granular materials[3] and on solidifying alloys[4,5].

Although modelling the behaviour of solidifying materials under deformation has already borrowed extensively from the field of granular mechanics[6,7], this pooling of resources can and should be pushed further. In particular, new approaches should be developed that recognize explicitly the discrete character of solidifying alloys[8]. Thus armed, we shall penetrate further into the mushy zones. ■

Christophe L. Martin is at the Laboratoire GPM2, CNRS, Institut National Polytechnique de Grenoble, 38402 Saint Martin d'Hères cedex, France.
e-mail: christophe.martin@inpg.fr

1. Gourlay, C. M. & Dahle, A. K. *Nature* **445,** 70–73 (2007).
2. Reynolds, O. *Phil. Mag.* **20,** 469–481 (1885).
3. Aste, T., Saadatfar, M. &. Senden, T. J. *Phys. Rev. E* **71,** 061302 (2005).
4. Ludwig, O., Dimichiel, M., Salvo, L., Suéry, M. & Falus, P. *Metall. Mater. Trans. A* **36,** 1515–1523 (2005).
5. Li, B., Brody, H. D. & Kazimirov, A. *Phys. Rev. E* **70,** 062602 (2004).
6. Flemings, M. C. *Metall. Trans. B* **22,** 269–293 (1991).
7. Ludwig, O., Drezet, J.-M., Martin, C. L. & Suéry, M. *Metall. Mater. Trans. A* **36,** 1525–1535 (2005).
8. Vernède, S., Jarry, P. & Rappaz, M. *Acta Mater.* **54,** 4023–4034 (2006).

EVOLUTIONARY BIOLOGY

Oxygen at life's boundaries

Peggy Baudouin-Cornu and Dominique Thomas

Proteins are made of amino acids. But amino acids are made of atoms. Exploration of this self-evident principle opens up fresh perspectives on the evolution of biological membranes and multicellular life.

For many microorganisms, one cell is adequate; for some plants and animals, billions are scarcely enough. But whatever the number, the cell is the fundamental unit of living matter, and is invariably delineated by a membrane — the plasma membrane — that is a selective barrier separating the inside from the outside. Some cells may also contain compartments, which are bounded by further membranes. Communication between intracellular compartments, or between cells and their environment, relies on transmembrane proteins that span the entire biological membrane. Using the unfamiliar prism of atomic rather than amino-acid composition, Acquisti *et al.*[1] show how their inspection of all the transmembrane proteins of 19 contemporary organisms tells us a lot about evolution. Their results appear on page 47 of this issue*.

Cells are divided into two large groups: eukaryotic, in which the DNA molecules are bounded by a nuclear membrane; and prokaryotic, which have no nuclear membrane. Prokaryotes are never found as complex, multicellular organisms. And whereas prokaryotes possess only simple intracellular compartments, or none at all, all eukaryotic cells contain compartments that are surrounded by two membranes. So understanding how and when compartmentalized cells appeared on Earth is one of the big questions in biology, as is understanding how and when multicellular eukaryotic organisms emerged millions of years later. Acquisti *et al.*[1] provide novel evidence of the absolute requirement of atmospheric oxygen (O_2) for these transitions to happen.

The 'oxygen revolution' stems from the first appearance, 3 billion years ago, of organisms releasing O_2 as a metabolic waste. This process led to a first great 'oxygenation event', 800 million years later, with a second one occurring one billion years ago. This second event is believed to have eventually fuelled the appearance of complex life-forms during the Cambrian explosion about 543 million years ago[2]. More recently, 425 million years ago, O_2 levels were a major factor in the progressive adaptation of aquatic arthropods and vertebrates to terrestrial life[3]. Accordingly, evolutionary analyses encompassing the past 2.3 billion years have revealed a correlation between increased organism complexity and the development of aerobic metabolism[4].

Two explanations have been given for this correlation, both invoking metabolic fitness. The first is that, compared with their anaerobic ancestors, oxygen-respiring cells are highly efficient energy-extracting machines: cells can use O_2 as an electron acceptor in respiration processes, and because of its high reduction potential, the maximum energy can then be released from nutritional resources. A second, complementary explanation stems from the observation that O_2 allows a thousand more metabolic reactions than can occur in anoxic conditions[5,6].

Acquisti *et al.*[1] now propose a third explanation, this time based on functional constraints. They argue that, in low O_2 conditions, it was impossible for cells to synthesize or maintain novel communication-related transmembrane proteins. Such proteins would be required for intracellular compartments to work together, a prerequisite to compartmentalization. Because evolution from unicellular to multicellular organisms requires efficient communication between cells, this evolutionary step was similarly hindered by insufficient levels of O_2.

*This article and the paper concerned[1] were published online on 20 December 2006.

NEUROBIOLOGY

Hit and miss

PHOTODISC/PHOTOLIBRARY

No matter how hard you practise a movement, you can never be entirely sure how it will turn out. Shouldn't the same action executed under the same conditions always produce the same result? Yet even professional darts players, throwing in a controlled indoor environment and standing a set distance from the board, can miss the bull's-eye.

Many theories of muscle control have assumed that such errors arise from variation generated during the movement — particularly 'noise' in the way that neurons pass instructions to the muscles at the neuromuscular junction. But Mark Churchland, Afsheen Afshar and Krishna Shenoy report that a large part of the problem could instead arise as the brain plans the action (*Neuron* **52,** 1085–1096; 2006).

They observed monkeys reaching for visual targets that appeared on a screen. When a target first appeared, it jittered slightly in place, and the animals were trained not to reach for it until it became stationary a half-second to a second later — allowing a period of preparation.

The authors recorded neural activity from the motor cortex and the premotor cortex, two brain regions involved in movement planning and execution. Comparing the monkeys' reaching movements with these recordings, they found that variations in the velocity of the reaches correlated with fluctuations in brain activity during the preparatory period — hundreds of milliseconds before the movement started.

So it seems that the execution of even a simple, well-practised task is limited by the brain's ability to plan the same movement over and over again. Indeed, Churchland and colleagues estimate that this constraint could account for at least half of the variability in the monkeys' movements.

Whether the fluctuations they observe actually arise in the premotor and motor cortex, or merely reflect variations elsewhere in the brain, is still an open question. And how variations in sensory input might affect the subsequent movement has yet to be fully explored. Finding the answers will have implications for our understanding of how the brain controls movement, and in the long term could have an impact on how movement disorders are treated.

Helen Dell

Surprisingly, Acquisti and colleagues' analyses suggest that the main distinctive feature of these novel transmembrane proteins is that they are enriched in oxygen atoms: in particular, their oxygen-rich external domains are longer than those of transmembrane proteins from uncompartmentalized cells.

How might levels of atmospheric oxygen have constrained the atomic composition of transmembrane proteins? Acquisti *et al.* propose two hypotheses. The first, inspired by stoichiometric ecology[7], is that, in the absence of O_2, building oxygen-rich amino acids would have been too demanding. However, there is no obvious evidence for such a metabolic limitation. According to computational analyses[5], the seven amino acids containing most oxygen atoms — D, E, Y, S, T, N and Q, in single-letter code — could all be synthesized by anoxic metabolisms. Moreover, among the 86 final reactions producing these amino acids, only 11 are specific to the oxic metabolism[5,8]. The authors thus favour the second hypothesis: that the reducing atmosphere found under low levels of O_2 would have damaged long, oxygen-rich protein domains, and made the synthesis of transmembrane proteins with long external parts impractical.

The work of Acquisti *et al.*[1] could be refined by correlating the oxygen content of transmembrane proteins with that of the compartments in which they are embedded. In particular, intracellular membranes delineating reducing compartments would be expected to contain more oxygen-poor proteins than does the plasma membrane. A candidate for such a study would be the membrane of mitochondria, the cell's energy-producing compartments. As a result of their continuous consumption of oxygen by respiration, mitochondria are the most anoxic compartments of oxygen-respiring cells.

Moreover, a striking difference between most eukaryotes and most prokaryotes is that respiration does not occur in the eukaryotic plasma membrane. As a consequence, O_2 is not consumed in the immediate proximity of eukaryotic plasma membranes, which thus exist in an oxygen-rich environment. Seen in the light of Acquisti and colleagues' paper, this may have been a factor in protecting their oxygen-rich transmembrane proteins. The O_2-driven emergence of multicellular organisms may therefore have required two major changes: accumulation of oxygen-rich proteins in the plasma membrane and confinement of respiration to intracellular compartments dedicated to that purpose.

It is striking that it has taken so long to make these simple observations on the elemental compositions of transmembrane proteins, and to formulate the resulting model of evolution[1]. The main reason may be that biologists usually regard proteins as chains of amino acids, or combinations of polypeptide domains, and ignore the fact that, in essence, proteins are arrangements of atoms. The elemental structure of biopolymers may well have been shaped by nutritional, physical or functional constraints[9], but the effects of these constraints usually remain hidden if one inspects only the amino-acid (or base-pair) compositions.

The work of Acquisti *et al.* is a welcome reminder that such constraints acted on subsets of proteins linked by function[10], cellular location[1] or metabolic role[11], as well as on the total protein content of a cell[12]. Structural biologists are no longer alone in keeping the atomic composition of their favourite proteins under close scrutiny — evolutionary biologists, too, will find this a fruitful pursuit. ■

Peggy Baudouin-Cornu is at the Service de Biochimie et Génétique Moléculaire, Département de Biologie Joliot-Curie, Direction des Sciences du Vivant, CEA/Saclay, 91191 Gif-sur-Yvette, France.
e-mail: peggy.baudouin@cea.fr
Dominique Thomas is at Cytomics Systems SA, 1 avenue de la Terrasse, 91190 Gif-sur-Yvette, France.
e-mail: dthomas@cytomics.fr

1. Acquisti, C., Kleffe, J. & Collins, S. *Nature* **445,** 47–52 (2007).
2. Knoll, A. H. *Life on a Young Planet: The First Three Billion Years of Evolution on Earth* (Princeton Univ. Press, 2003).
3. Ward, P., Labandeira, C., Laurin, M. & Berner, R. A. *Proc. Natl Acad. Sci. USA* **103,** 16818–16822 (2006).
4. Hedges, S. B., Blair, J. E., Venturi, M. L. & Shoe, J. L. *BMC Evol. Biol.* **4,** 2 (2004).
5. Raymond, J. & Segrè, D. *Science* **311,** 1764–1767 (2006).
6. Falkowski, P. G. *Science* **311,** 1724–1725 (2006).
7. Sterner, R. W. & Elser, J. J. *Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere* (Princeton Univ. Press, 2002).
8. http://prelude.bu.edu/O2
9. Baudouin-Cornu, P. & Bragg, J. G. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (eds Jorde, L., Little, P., Dunn, M. & Subramanian, S.) Section 3.3 (doi:10.1002/047001153X.g303318) (Wiley, New York, 2006).
10. Mazel, D. & Marlière, P. *Nature* **341,** 245–248 (1989).
11. Baudouin-Cornu, P., Surdin-Kerjan, Y., Marlière, P. & Thomas, D. *Science* **293,** 297–300 (2001).
12. Elser, J. J., Fagan, W. F., Subramanian, S. & Kumar, S. *Mol. Biol. Evol.* **23,** 1946–1951 (2006).

COSMOLOGY

Ripples of early starlight

Craig J. Hogan

After all known sources are accounted for, puffy blobs of infrared light persist on deep-field telescope images. Evidence is mounting that these could be the signatures of stars in early 'protogalaxies'.

About a year ago, Kashlinsky *et al.* found evidence for fluctuations in background infrared light from far-off cosmological sources larger than would be expected from unresolved galaxies in known populations[1,2]. In two papers to be published in *The Astrophysical Journal*, the same authors now confirm the effect in different and larger sections of the sky[3,4]. They argue that the most plausible interpretation is fluctuations in the light emitted by an early generation of stars. If so, this is the oldest starlight yet detected, and dates from long before modern giant galaxies formed.

In the conventional model of early cosmic evolution, small primordial concentrations in the density of unseen 'dark matter' had, about 100 million years after the Big Bang, finally grown large enough to cause the gravitational collapse of gas towards them. This process created pockets of gas dense enough to collapse of their own accord and form the first stars. By the time the Universe was about 200 million years old, stellar burning in small, pregalactic systems was widespread. After this time, stars formed in successively larger, hierarchically assembled systems. After a few billion years, this evolution culminated in the formation of the giant galaxies that we see all over the Universe today.

Detailed, direct observations of pregalactic starlight from the Universe's first billion years would allow astronomers to probe many aspects of the early evolution of the cosmos that are obscure or untested. These include the formation of the first clumps, or 'minihaloes', of dark matter; the collapse of the first clouds of gas on small scales; the formation and explosion of the first stars, and the end of the cosmic 'dark ages'; the chemical enrichment and ionization of the early cosmos; and the energy feedback that apparently kept most gas from ever forming stars.

Kashlinsky and colleagues' latest analysis[3] was performed, as before[1], on deep-field exposures made by NASA's Spitzer Space Telescope. After identifiable sources are masked out and digitally cleaned from the data, a background of fuzzy blobs remains (Fig. 1) that is not consistent with randomly distributed 'shot' noise from unresolved sources. The images show additional fluctuations on scales of about 1 to 10 arcminutes (1/60th to 1/6th of a degree), this upper limit being set by the size of the surveyed fields.

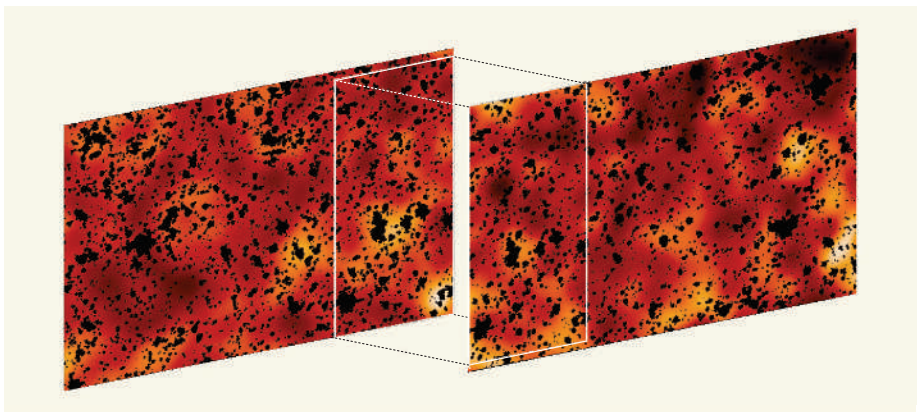
The authors' favoured interpretation[4] of this effect is that the blobs represent a fractional fluctuation of a few to about ten per cent in a rather intense background of starlight originating in the first 1.2 billion years of the Universe's existence. This equates to sources at redshifts (z) of between about 5 and 20 (where $1 + z$ is the factor by which the Universe has expanded since the source emitted the light). Individual, collapsed star-forming systems associated with early protogalaxies are small and numerous, so their random shot noise is small, and in the absence of other perturbations their light should be relatively smoothly distributed. But standard cosmology predicts that these systems are clustered non-randomly owing to the larger-scale and smaller-amplitude primordial fluctuations in matter density that eventually become giant galaxies and groups of galaxies. Summed along the line of sight, these fluctuations give rise to the detected blobs. (The brighter, uniform, non-fluctuating component of the background is not directly detectable in these data, because it cannot be distinguished from other sources of noise and emission.)

That this starlight should appear as puffy blobs at infrared wavelengths is entirely expected[5]. In everyday life, distant things — whose light has travelled further to reach us — look smaller. But looking back to redshifts greater than about 1.6, systems in our expanding Universe start appearing larger, not smaller. This odd effect is caused by the expansion itself, which not only shifts the visible spectrum of starlight to longer, infrared frequencies, but also means that older objects at higher redshift were 'closer' at the time they emitted the light. Thus, an object has about the same angular size at redshift 10 as it does at redshift 0.3. At redshift 10, the angular size of the observed blobs, about 100 arcseconds, corresponds to a size of some 1.3 million light years. This size is typical of the dark-matter haloes that eventually collapse into giant galaxies and groups of galaxies (but, at the time the light was emitted, were yet to collapse).

At present, the data are not specific enough to allow us to tell where the light is coming from, but a high-redshift interpretation works if the emitting stars are hot, massive and bright. Alternatively, the light might be coming from lower redshifts, but in this case its smooth distribution, and the fact that more starlight has not been seen at optical wavelengths, are puzzling.

Either way, the much more powerful telescopes on the way, especially NASA's James Webb Space Telescope, which is scheduled for launch within the next decade, will have plenty to study. This telescope's much larger primary mirror will allow it to resolve the small regions of protogalactic star formation, perhaps find spectroscopic clues as to the redshift of the light, and maybe even detect supernova explosions as the first stars die in the regions surveyed. The evidence seems to be growing that there is more than just a dark sky between the galaxies: ripples of infrared starlight glow in every direction. ■

Craig J. Hogan is in the Departments of Astronomy and of Physics, University of Washington, Box 351580, Seattle, Washington 98195-1580, USA.
e-mail: hogan@u.washington.edu



A. KASHLINSKY/NASA

Figure 1 | Infrared map. These images of a portion of the sky in the Hubble Deep Field North are based on data from NASA's Spitzer Space Telescope at two different infrared wavelengths, 3.6 μm and 4.5 μm, including an overlapping region as shown. The black pixels are close to bright sources and are masked off. The remaining pixels are cleaned of point sources, leaving extended fuzzy blobs glowing in the background that show up in both bands. These could be fluctuations, due to nascent cosmic structure, in a bright pregalactic infrared background.

1. Kashlinsky, A., Arendt, R. G., Mather, J. & Moseley, S. H. *Nature* **438,** 45–50 (2005).
2. Ellis, R. *Nature* **438,** 39 (2005).
3. Kashlinsky, A., Arendt, R. G., Mather, J. & Moseley, S. H. *Astrophys. J.* (in the press); preprint available at www.arxiv.org/astro-ph/0612445 (2006).
4. Kashlinsky, A., Arendt, R. G., Mather, J. & Moseley, S. H. *Astrophys. J.* (in the press); preprint available at www.arxiv.org/astro-ph/0612447 (2006).
5. Bond, J. R., Carr, B. J. & Hogan, C. J. *Astrophys. J.* **306,** 428 (1986).

REVIEWS

Light in tiny holes

C. Genet[1] & T. W. Ebbesen[1]

The presence of tiny holes in an opaque metal film, with sizes smaller than the wavelength of incident light, leads to a wide variety of unexpected optical properties such as strongly enhanced transmission of light through the holes and wavelength filtering. These intriguing effects are now known to be due to the interaction of the light with electronic resonances in the surface of the metal film, and they can be controlled by adjusting the size and geometry of the holes. This knowledge is opening up exciting new opportunities in applications ranging from subwavelength optics and optoelectronics to chemical sensing and biophysics.

A hole in a screen is probably the simplest optical element possible, and was an object of curiosity and technological application long before it was scientifically analysed. A pinhole was at the heart of the camera obscura used by the Flemish painters in the sixteenth century to project an image (albeit upside down) onto their canvases. It was in the middle of the seventeenth century that Grimaldi first described diffraction from a circular aperture[1], contributing to the foundation of classical optics. Despite their apparent simplicity and although they were much larger than the wavelength of light, such apertures remained the object of scientific study and debates for centuries thereafter, as an accurate description and experimental characterization of their optics turned out to be extremely difficult.

In the twentieth century, the interest naturally shifted to subwavelength holes as the technology evolved towards longer wavelengths of the electromagnetic spectrum. With the rising importance of microwave technology in the war effort of the 1940s, Bethe treated the diffractive properties of an idealized subwavelength hole, that is, a hole in a perfectly conducting metal screen of zero thickness[2] (Box 1). His predictions, notably that the optical transmission would be very weak, became the reference for issues associated with the miniaturization of optical elements and the development of modern characterization tools beyond the diffraction limit, such as the scanning near-field optical microscope (SNOM), which typically has a small aperture in the metal-coated tip as the probing element[3].

In this context, the report of the extraordinary transmission phenomenon through arrays of subwavelength holes milled in an opaque metal screen[4] generated considerable interest because it showed that orders of magnitude more light than Bethe's prediction could be transmitted through the holes. This has since stimulated much fundamental research and promoted subwavelength apertures as a core element of new optical devices. Central to this phenomenon is the role of surface waves such as surface plasmons (SP), which are essentially electromagnetic waves trapped at a metallic surface through their interaction with the free electrons of the metal[5,6] (Box 2). This combination of surface waves and subwavelength apertures is what distinguishes the enhanced transmission phenomenon from the idealized Bethe treatment and gives rise to the enhancement. Moreover, modern nanofabrication techniques allow us to tailor the dynamics of this combination by structuring the surface at the subwavelength scale. This opens up a wealth of possibilities and applications from chemical sensors to atom optics.

We will review here the present understanding of the transmission through subwavelength apertures in metal screens, starting for the sake of clarity with simple isolated holes and ending with arrays. As we will see, SPs play an essential role at optical wavelengths in all the considered structures. Applications such as tracking single molecule fluorescence in biology, enhanced vibrational spectroscopy of molecular monolayers and ultrafast photodetectors for optoelectronics illustrate the broad implications for science and technology.

Single apertures

Figure 1a shows a single hole milled in a free-standing Ag film, characterized by both the diameter of the hole and its depth. When Bethe considered such a system, he idealized the structure by assuming that the film was infinitely thin and that the metal was a perfect conductor. With these assumptions, he derived a very simple expression for the transmission efficiency η_B (normalized to the aperture area)[2]:

$$\eta_B = 64(kr)^4/27\pi^2 \quad (1)$$

where $k = 2\pi/\lambda$ is the norm of the wavevector of the incoming light of wavelength λ, and r is the radius of the hole. It is immediately apparent that η_B scales as $(r/\lambda)^4$ and that therefore we would expect the optical transmission to drop rapidly as λ becomes larger than r, as shown in Box 1. In addition, the transmission efficiency is further attenuated exponentially if the real depth of the hole is taken into account[7]. This exponential dependence reflects the fact that the light cannot propagate through the hole if $\lambda > 4r$, whereupon the transmission becomes a tunnelling process. The cutoff condition $\lambda > 4r$ is of course a first approximation and in real situations the cutoff occurs at longer wavelengths when the finite conductivity is taken into account[8] (see Box 1).

Bethe also predicted that the light would diffract as it emerges from the hole in an angular pattern that depends on the orientation relative to the polarization of the incident light[2]. If the diffraction pattern is scanned along the direction of the incoming polarization the intensity should be constant (like a spherical wave in a plane) while in the perpendicular direction, the intensity decreases with increasing angle (the angular dependence is a $\cos^2\theta$ function, typical of a dipole emission pattern).

The increasing use of SNOM and interest in the extraordinary transmission phenomenon have stimulated experimental[9–11] and theoretical[12–16] studies, the results of which challenge Bethe's predictions. In particular, it has become possible to measure the transmission and diffraction from a single subwavelength aperture in a metallic film at optical wavelengths[9–11]. Angular measurements at

[1]ISIS, Université Louis Pasteur and CNRS (UMR7006), 8 allée G. Monge, 67000 Strasbourg, France.

the exit of subwavelength apertures have revealed that the light diffracts less than expected[9,10]. Similarly, the transmitted light can have unexpected features[10]. The simple circular aperture of Fig. 1a has a transmission spectrum with a peak as shown in Fig. 1b not predicted by equation (1) or by other conventional theories[2,7]. Similar measurements can be made on a rectangular hole (Fig. 1c) where the spectrum becomes sensitive to the incident light polarization as can be seen in Fig. 1d. The appearance of such resonant peaks can be understood as the excitation of SP modes at the edges of the hole, a type known as localized SP that has been confirmed by theoretical studies[14]. By aligning the incoming polarization on either the short or long axis of the rectangular hole, we can selectively excite the corresponding localized SPs (Fig. 1d). Such behaviour is very reminiscent of elongated metal particles, the colours of which are also determined by localized SPs. Whereas the localized SP modes are defined by the lateral dimensions of the aperture, theoretical studies have shown that in addition to such SP modes[14] other resonant modes defined along the depth of the hole might also be present and contribute to the transmission signal[12]. Further experimental studies on this issue at optical wavelengths are necessary.

Bethe's theory describes the transmission as a smooth decreasing function of the wavelength, as given by equation (1) and shown in Box 1, whereas the experiments discussed above reveal the presence of a resonance superimposed on a smooth background, thus providing an enhancement at the resonant wavelengths. In all the structures presented in this review, it is always the presence of some type of resonance that leads to transmission enhancement. This reveals yet again that Bethe's theory is too idealized to treat situations where surface modes are involved and where propagating or evanescent modes can additionally be excited inside the hollow aperture[12], thereby significantly underestimating the transmission efficiency. We define the transmission as being extraordinary when it is so enhanced that the transmission efficiency η is larger than 1, in other words when the flux of photons per unit area emerging from the hole is larger than the incident flux per unit area. As we shall see in the next sections, η can be much larger than one for certain aperture structures under appropriate conditions.

For experimental reasons, it is very difficult to quantify η for a single hole. As was pointed out above, the emission pattern from a singe aperture in a real metal is not isotropic and therefore the absolute transmission can only be determined by measuring the absolute intensity over all angles and then summing the data. This remains an experimental challenge. As we shall see in the section on optimizing

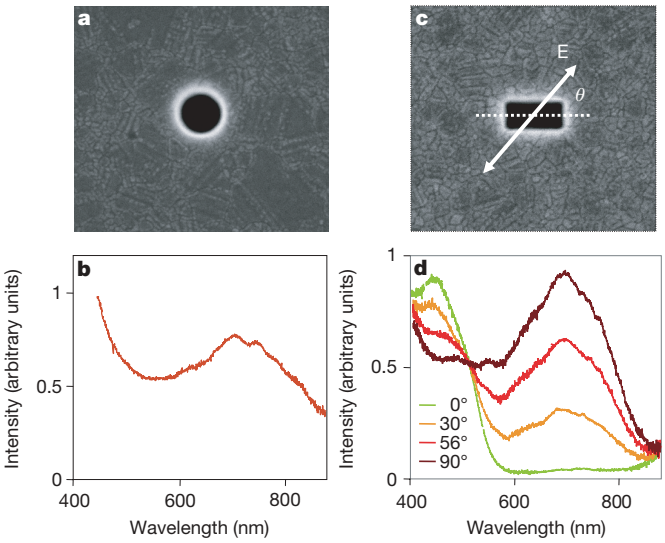
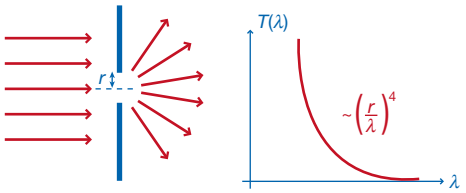


Figure 1 | Optical transmission properties of single holes in metal films. The holes were milled in suspended optically thick Ag films illuminated with white light. **a**, A circular aperture and **b**, its transmission spectrum for a 270 nm diameter in a 200-nm-thick film. **c**, A rectangular aperture and **d**, its transmission spectrum as a function of the polarization angle θ for the following geometrical parameters: 210 nm × 310 nm, film thickness 700 nm. Figure adapted from ref. 10, with permission.

Box 1 | Light transmission through apertures

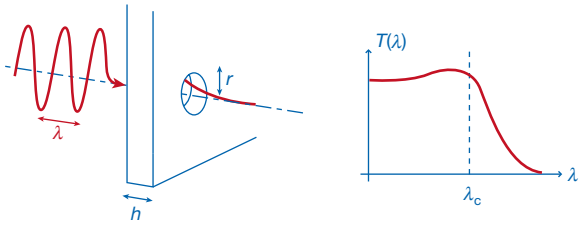
When light scatters through apertures, it diffracts at the edges. In the subwavelength regime, Bethe was able to give a theoretical description of the diffraction of light at a given wavelength λ through a circular hole of radius $r << \lambda$ in the idealized situation of an infinitely thin and perfect metal sheet. He has shown that the transmission $T(\lambda)$ scales uniformly with the ratio of r to λ to the power of four, as described in equation (1) and schematically shown below in Box 1 Fig. 1.



Box 1 Figure 1 | Diffraction and typical transmission spectrum of visible light through a subwavelength hole in an infinitely thin perfect metal film.

However, a real aperture is characterized by a depth and therefore has waveguide properties. The transmission of light through such a guide is very different from the propagation of light in empty space. The confined space of the waveguide essentially modifies the dispersion relation of the electromagnetic field. The lateral dimensions of the waveguide define the wavelength at which light can no longer propagate through the aperture. This wavelength is known as the cutoff wavelength λ_c. When the incident wavelength $\lambda > \lambda_c$ the transmission is exponentially small, characterizing the non-propagating regime as shown in Box 1 Fig. 2. With real metals, the cutoff wavelength cannot be sharply defined because one goes continuously from propagative to evanescent regime as the wavelength increases.

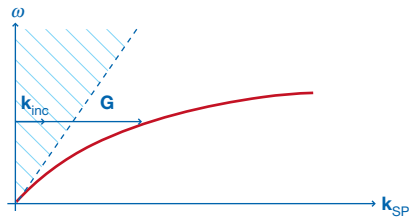
There is a straightforward relationship between the cross-section of the waveguide and λ_c. However, one should take into account that λ_c for an aperture in a real metal is increased by taking the skin-depth into account, reflecting the penetration of the electromagnetic field inside the walls of the metal waveguide. It is possible to control and even to eliminate cutoff wavelengths even when the lateral dimensions are much smaller than λ, by playing with more complex geometries. While simple apertures are always characterized by the existence of cutoff wavelengths, an annular hole, for example, which resembles a coaxial cable, has no cutoff wavelength and is always propagating. The polarization of the incident light is also an important parameter, and with non-cylindrical waveguides, the transmission can be made extremely polarization sensitive. A striking illustration is provided by a slit. Here, for incident polarization parallel to the long axis, the transmission can be made subwavelength, as soon as the short dimension of the rectangle is smaller than λ. However, for the perpendicular polarization, no matter how narrow the guide is, the light always propagates through it. This allows for many possibilities in the choice of geometry depending on the application.



Box 1 Figure 2 | A cylindrical waveguide with a radius *r* much smaller than the wavelength λ of the incident electromagnetic field milled in a metal film of thickness *h*. The exponentially decreasing tail represents the attenuation of the subwavelength regime. A transmission spectrum can reveal the different propagating and evanescent regimes.

Box 2 | Coupling to SPs

At the interface separating a dielectric with a permittivity ε_d and a metal with a permittivity ε_m, SPs can be resonantly excited by the coupling between free surface charges of the metal and the incident electromagnetic field. Such a mode is characterized by a surface wave vector that obeys the following dispersion relation:



Box 2 Figure 1 | SP dispersion relation. The dotted line corresponds to the light line. The hatched sector of propagating waves does not overlap with the evanescent sector below the light line that fully contains the SP dispersion relation. $\mathbf{k}_{inc}$ is the transverse component of the incident wave vector and **G** corresponds to the momentum needed to couple to the SP mode in the evanescent sector.

$$k_{SP} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m \varepsilon_d}{\varepsilon_m + \varepsilon_d}}$$

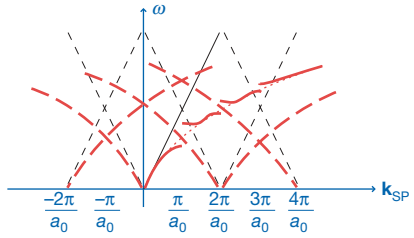
Here, ω is the pulsation of the electromagnetic field and c the velocity of light in vacuum. Provided that the real part of ε_m is smaller than $-\varepsilon_d$, this wave vector has positive real and imaginary parts. The latter corresponds to the propagation length of the surface wave before it is damped inside the metal, and can be tens of micrometres at the smooth surface of noble metals, such as Au or Ag, at optical wavelengths. The real part of k_{SP} is plotted in Box 2 Fig. 1. It is always below the light line that separates free-space photons from evanescent ones. This implies immediately that such a mode is evanescent and therefore cannot be excited directly by freely propagating light. A given additional momentum **G** is needed to go from the propagating sector where the wave vector $\mathbf{k}_{inc}$ of the incident light falls to the evanescent one where SP modes exist. This is expressed in the simple resonance condition $\mathbf{k}_{SP} = \mathbf{k}_{inc} + \mathbf{G}$, which is a function of the incident pulsation and incident angle θ.

One way to provide the missing momentum **G** necessary for coupling incoming light to SPs is to use a periodic array. In one dimension for instance, it can be shown that **G** is related to multiples of $2\pi/a_0$ where a_0 is the period of the structure. This is the origin of the optical resonant behaviour of the array, because only when:

$$k_{SP} = k_0 \sin\theta + i \times \frac{2\pi}{a_0}$$

does light couple to SPs (i is an integer). The electromagnetic wave is then trapped momentarily on the surface, giving rise to the transmission peaks. The array generates a complex band structure, as schematically shown in Box 2 Fig. 2. At every multiple of π/a_0 (Brillouin zone edges), SPs are back-reflected so strongly that they cannot propagate any more. Bandgaps appear in the SP dispersion relation, corresponding to stationary waves and high field enhancements.

It should be noted that, when illuminated, non-periodic structures such as single holes, sharp edges, particles and so on can generate localized SP modes. This is possible when the dimensions of the defects are smaller than the wavelength of the incident field, generating a broad spectrum of **G** vectors (stemming from the spatial Fourier spectrum of the particular defect) in which a solution to the coupling condition $\mathbf{k}_{SP} = \mathbf{k}_{inc} + \mathbf{G}$ can be found. The coupling efficiency is dependent on the particular profile of the defect.



Box 2 Figure 2 | SP band structure on a periodic array.

subwavelength apertures below, the best transmission signals are obtained in noble metals such Au and Ag. To obtain detectable resonances at visible wavelengths, the dimensions of the holes should be of the order of 150 to 300 nm and the films not much thicker than 200 to 300 nm.

Small apertures are routinely used in SNOM tips to explore and to map with subwavelength resolution the electromagnetic field in the immediate vicinity of a surface[3]. More recently, tiny apertures have been implemented in fluorescence correlation spectroscopy[17–19], a powerful technique for the study of the diffusion and reaction of single fluorescent biomolecules in which the information is derived from the analysis of the statistical fluctuations of individual molecules as they move through a small volume. Traditionally, the volume is defined by the focal point of a laser beam, that is, about 1 μm^3, which puts a limit on the upper concentration that can be used while still observing statistical fluctuations. By using small apertures in metal (see Fig. 2)[18,19], the analysed volume has been reduced by a factor of 1,000, allowing one to study molecular events at nearly millimolar concentrations—closer to biological conditions. In addition, such structures give rise to other benefits: the localized SP fields increase the excitation rate of the molecules in its vicinity[10,14,19], the emission pattern is potentially directional[9,10] and the branching ratios from the fluorescent state are affected[19]. All these can lead to an increase in the detected signal, rendering fluorescence correlation spectroscopy ever more useful as a tool for biology.

Single apertures surrounded by periodic corrugations

With modern nanofabrication techniques it is possible to modify the optical properties of a single aperture by sculpting the surrounding material at the scale of the wavelength[20–28]. Such modifications give rise to much higher transmission than single holes at selected wavelengths and in addition, novel lensing effects including beaming can be induced by texturing the output surface of the aperture—as discussed next.

When a single aperture is surrounded by circular corrugations as shown in Fig. 3a, the periodic structure acts like an antenna to couple the incident light into SPs at a given λ. As a consequence the electromagnetic fields at the surface become intense above the aperture, resulting in very high transmission efficiencies and a well-defined spectrum (Fig. 3a). Here the resonant wavelength is mainly determined by the periodicity of the grooves, which provides the necessary momentum and energy-matching conditions (as explained in Box 2). The resonance is, however, slightly more red-shifted than the period owing to the interaction with the light directly transmitted through the hole. This should be considered in tuning the structure to be bright at a desired wavelength. When such a structure is milled in a metal like Ag, the value of η can be much larger than one[21]. Again, absolute quantification is difficult, but compared to a bare single hole of the same dimensions the transmission gain can be an order of magnitude at resonant wavelengths[20]. This, as we shall see below, has important applications.

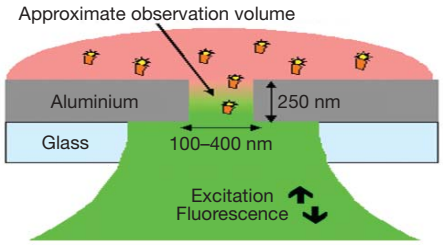


Figure 2 | Schematic diagram of the fluorescence correlation spectroscopy in a single hole. The fluorescence of individual molecules is collected as they diffuse through the observation volume defined by the hole in the metal film. The fluorescence is collected from the same side as the incident excitation. Courtesy of J. Wenger.

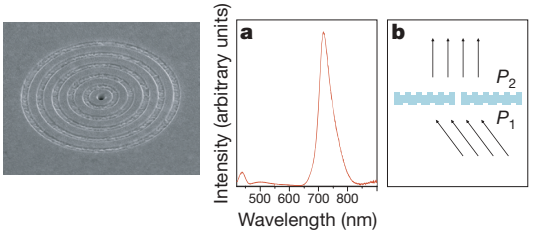


Figure 3 | Optical properties of single apertures surrounded by periodic corrugations. **a**, Transmission spectrum of a single hole surrounded by periodic corrugations (left) prepared by focused ion beam (hole diameter 300 nm, period 650 nm). **b**, Schematic illustration of redirecting beam by single-slit aperture surrounded by grooves of different periodicity on the input (P_1) and output (P_2) surfaces.

If the output surface surrounding the aperture is also corrugated, a surprisingly narrow beam can be generated, having a divergence of less than a few degrees[20], which is far smaller than that of the single apertures discussed earlier. This is because the light emerging from the hole couples to the periodic structure of the exit surface and to the modes existing in the grooves—which in turn scatter the surface waves into freely propagating light[22–24]. This then interferes with the light that has travelled directly through the hole generating the focused beam. A variant of the double-sided bull's-eye structure is a slit with parallel grooves on both sides of the film, which in addition disperses light spatially according to wavelength[20,28]. Such double-sided structures act as a novel kind of optical element[20,23–25]. They can have a focal plane like a lens but at the same time have other unusual features. For instance, by having grooves with different periods on either side of the film next to a slit, the direction of the output beam can be made independent of the input beam, unlike conventional lenses or gratings, suggesting many practical applications (Fig. 3b). This ability to redirect the beam stems from the way input and output corrugations act like two separate independent gratings connected by a pinhole.

The antenna capacity of the corrugations to concentrate the photons at the tiny central aperture also opens up other technological possibilities such as a bright subwavelength spot for ultradense optical-data storage and nonlinear phenomena[29–31]. Of paramount importance to modern optical telecommunication are photodetectors that can translate an optical signal into an electrical one and thereby convert the flow of information being carried through the telecom network into a displayable signal on the screen. Such photodetectors must therefore be as fast as possible to handle the large amount of data flow. Typically the operating speed of a photodetector scales inversely with the size of its photoelectrical element, but the size cannot be made too small because then it would no longer collect enough photons. To circumvent this problem, an ultrafast photodetector has been realized that elegantly combines a very small photoelectrical element with a bull's-eye antenna structure (shown in Fig. 4) that collects and concentrates the incoming photons[32]. This combination illustrates well the potential benefit of plasmonic devices for optoelectronics.

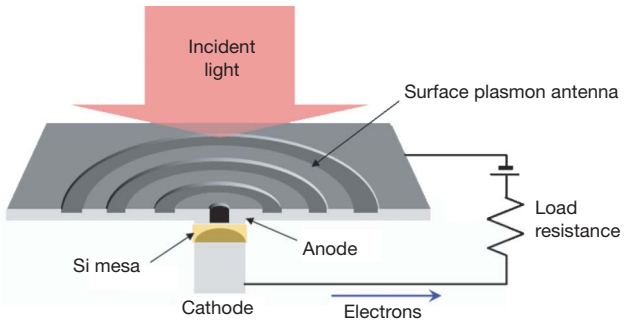


Figure 4 | Ultrafast miniature photodetector. This device consists of a small Si photoelectric element and a SP antenna (reproduced from ref. 32 with permission). The incoming light is harvested by the periodic structure surrounding the central hole, which then transmits it to the underlying photodetector.

Hole arrays

Periodic arrays of holes in an opaque metal film have so far been the structures that have found the most applications owing to the simplicity with which their spectral properties can be tuned and scaled. Among other things, they can act as filters for which the transmitted colour can be selected by merely adjusting the period. As we saw in the previous section (and Box 2), periodic metallic structures can convert light into SPs by providing the necessary momentum conservation for the coupling process. It is therefore not surprising that periodic arrays of holes such as those shown in Fig. 5 can give rise to the extraordinary transmission phenomenon[4] where the transmission spectrum contains a set of peaks with η larger than one even when the individual holes are so small that they do not allow propagation of light (Fig. 5). Hole arrays have been characterized in great detail both theoretically[33–59] and experimentally[58–79]. As in the case of single holes surrounded by periodic grooves[22], the process can be divided into three steps: the coupling of light to SPs on the incident surface, transmission through the holes to the second surface and then re-emission from the second surface. At the peak transmissions, standing SP waves are formed on the surface (Box 2). The intensity of SP electromagnetic fields above each hole compensates for the otherwise inefficient transmission through each individual hole (Box 1).

If we apply the momentum-matching conditions discussed in Box 2 to a two-dimensional triangular array shown in Fig. 5, we can show that the peak positions λ_{max} at normal incidence are given in a first approximation by:

$$\lambda_{max} = \frac{P}{\sqrt{\frac{4}{3}(i^2+ij+j^2)}}\sqrt{\frac{\varepsilon_m\varepsilon_d}{\varepsilon_m+\varepsilon_d}} \qquad (2)$$

where P is the period of the array, ε_m and ε_d are respectively the dielectric constants of the metal and the dielectric material in contact with the metal and i, j are the scattering orders of the array. Because equation (2) does not take into account the presence of the holes and the associated scattering losses, it neglects the interference that gives rise to a resonance shift[42,43]. As a consequence, it predicts peak positions at wavelengths slightly shorter than those observed experimentally, as can be seen in Fig. 5.

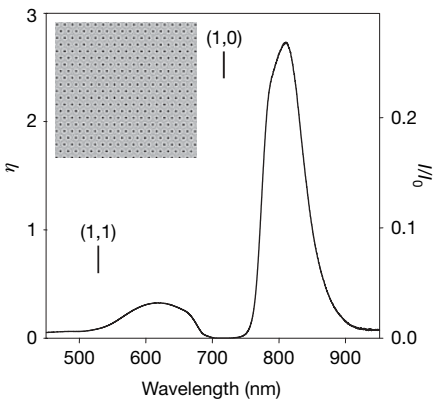


Figure 5 | Transmission spectrum of hole arrays. The triangular hole array was milled in a 225-nm-thick Au film on a glass substrate with an index-matching liquid on the air side (hole diameter 170 nm, period 520 nm). The transmission spectrum is measured at normal incidence using collimated white light. The inset shows the image of the actual array. I/I_o is the absolute transmission of the array and η is the same transmission but normalized to the area occupied by the holes.

Implicit in the resonance conditions defined by equations such as equation (2) are the symmetry relations of the array. Therefore the SPs generated in the array will propagate along defined symmetry axes with their own polarization depending on the (i, j) number of the mode. This results in a rich polarization behaviour that can be revealed in particular under focused light illumination[79].

We emphasize that both surfaces on either side of the holes can sustain SP modes offset from each other by the difference in ε_d of the material in immediate contact with the metal surface (typically glass and air), as predicted by equation (2). Hence, the transmission spectrum of asymmetric structures contains two sets of peaks, each set belonging to one of the surfaces. In many applications, hole arrays of a finite size are used for practical reasons. If the arrays contain small numbers of holes then the periodicity is not well defined and the contribution from the edges becomes significant, changing the spectrum and leading to unusual re-emission patterns[33].

One interesting feature of hole arrays is the fact that each hole on the output surface acts like a new point source for the light. Therefore, if a plane wave (that is, a collimated beam) impinges on the input surface, then a plane wave is reconstructed through classical interference as the light travels away from the output surface. Naturally, because the array is also a grating, the transmission gives rise to different diffraction orders depending on the wavelength to period ratio. For the longest-wavelength (1,0) peak shown in Fig. 5, only the 0th order diffraction is formed, because $\lambda > P$. When $\lambda < P$, higher diffraction orders gradually appear as the wavelength becomes shorter.

The shape and dimensions of the holes in an array do influence its transmission spectrum[64,65]. For instance, in the case of non-propagative apertures, switching from circular to rectangular holes changes the spectrum as a result of the simultaneous change in both the localized SP mode associated with each individual hole and the cutoff wavelength (the wavelength above which the aperture no longer allows light propagation; see Box 1). Nevertheless, the spectrum is dominated by the SP modes because of the periodicity of the array[65]. If the transmission peak falls below the cutoff, its intensity drops exponentially with the depth of the hole and hence the film thickness (or hole depth) is a critical parameter in these structures[33]. It should be noted that arrays of slits have more complex transmission spectra than do hole arrays because the slits can be made propagative under the appropriate polarization (Box 1). As a consequence, the transmission spectra typically contain the signature of both cavity modes in the slits (often at wavelengths that equal twice the slit depth divided by an integer) and SP modes, owing to the slit periodicity[35,48–54,61].

Hole arrays have many applications, from optical elements to sensors for chemistry and biology. For instance, the array acts like a tunable filter because the wavelength selectivity of the array transmission can be adjusted simply by changing the period, as predicted by equation (2) and illustrated in Fig. 6. The letters 'hν' are obtained by fabricating a periodic dimple array in which some of the dimples are milled through to form holes, which in turn reveal the spectral signature of the array.

The combination of the large electromagnetic fields generated by the SPs on the hole arrays, their sensitivity to the dielectric medium in contact with the surface (equation (2)) and the simplicity of the arrays have spurred efforts to use them to detect molecules and to enhance spectroscopic signals (fluorescence, Raman and so on)[80–89]. In this perspective, the enhanced infrared molecular vibrational spectroscopy exemplifies well the usefulness of the hole arrays for chemistry and biology[80]. Arrays of square holes in a Ni film with a periodicity tuned to the infrared were prepared and modified with Cu oxide to induce the catalytic transformation of methanol to formaldehyde on the surface. When such an array with a single adsorbed molecular layer on surface is placed in a Fourier-transform infrared apparatus, the infrared vibrational absorption spectrum (Fig. 7) that is extracted is at least 100 times stronger than if the apparatus had probed a molecular layer on an inert dielectric substrate[80,87]. The large signal enhancement is due to the fact that when the light is trapped momentarily on the surface in terms of SPs, its interaction time with the molecules increases and therefore so does the probability of the absorption. Note that absorption enhancement of electronic transitions is also observed in the visible part of the spectrum, although the enhancement factor is then only a factor of ten, owing to the shorter SP lifetime on the surface at optical wavelengths[90]. Needless to say, such results are extremely promising for studying molecular monolayers and surface reactions by static or time-resolved spectroscopy because this method provides a much better signal-to-noise ratio and is relatively easy to implement.

Other considerations for optimizing apertures

So far we have discussed mainly the broad features associated with tiny holes in opaque metal films. The unique properties of these holes being related to the presence of SPs, they are in turn very much dependent on geometric factors and the properties of the metal.

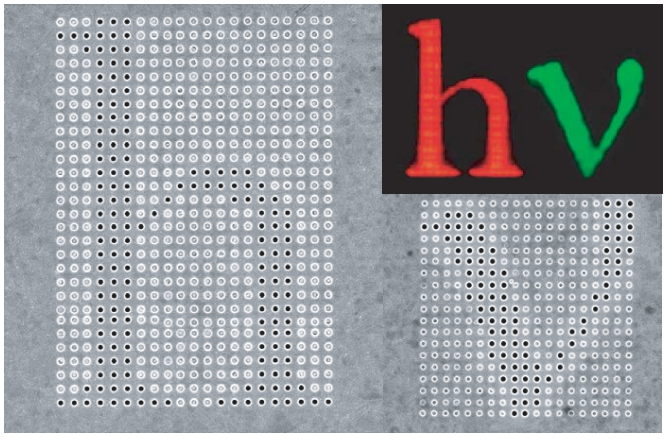


Figure 6 | Holes in a dimple array generating the letters 'hν' in transmission. An array of dimples is prepared by focused-ion-beam milling an Ag film. Some of the dimples are milled through to the other side so that light can be transmitted. When this structure is illuminated with white light, the transmitted colour is determined by the period of the array. In this case the periods were chosen to be 550 and 450 nm respectively to achieve the red and green colours.

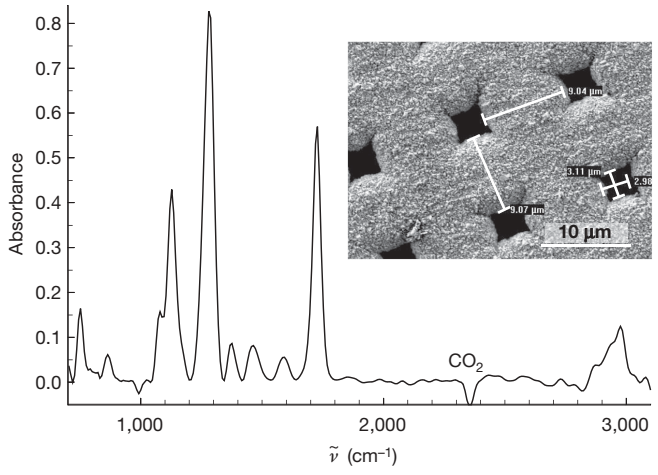


Figure 7 | Infrared enhanced vibrational spectra. Vibrational absorption spectrum of formaldehyde (CH_2O) monolayer adsorbed on a Ni hole array covered with Cu oxide (adapted from ref. 81 with permission). Note that absorbance of 0.8 implies that 84% of the incident light is absorbed by the molecular monolayer. The Ni hole array here acts like the antenna, trapping the light momentarily on the surface and therefore increasing the likelihood of absorption by the CH_2O.

The choice of the metal depends on the wavelength to be used because the dielectric constants of metals are wavelength dependent. Ideally, the dielectric constant of the metal should have a high absolute value for the real part and a small imaginary part that determines the absorption into the metal. This combination gives rise to high SP fields at the surface and minimizes the losses. Therefore Ag is ideally suited to obtain high transmission in the visible part of the spectrum, while above 600 nm Au is even better because it suffers little oxidation. In the infrared, metals such as Ni or Cu can also be used.

Interestingly, high transmittivity through structures similar to those described above but scaled to the microwave region of the spectrum have also been reported where SPs are not considered to exist[91–95]. This is explained by the fact that when a metal surface is corrugated, it is the effective dielectric constant that is modified rather than the bulk metal[37] (this is similar to the way the wetting properties of a material are changed by nanostructuring). As a consequence, SP-like surface waves (also known as 'spoof plasmons') are formed, which enhances the transmittivity[94,95]. More generally, it is essential to trap the electromagnetic wave in the vicinity of the aperture to observe enhanced transmission and its related phenomena. We therefore expect that surface waves such as surface phonon polaritons can also be used. There has been some discussion on whether SPs are involved in the optical transmission of aperture structures but a recent analysis has confirmed the key role of SPs[96,97], in agreement with the vast majority of studies.

The geometrical factors that influence the optical properties of the holes are numerous: symmetry of the structure, the aspect ratios and shape of the holes, aperture area, profile of the corrugations and so forth. These variables determine the electromagnetic field distribution on the surface, the propagation dynamics of the surface waves and their scattering efficiencies, and, the in-plane and out-of-plane coupling to light. The depth of the holes (or thickness of the metal film) is important for several reasons. If the films are too thin, they are partially transparent to the incident light and no holes are necessary to achieve significant transmission, especially if the surfaces are in addition resonantly corrugated[98,99]. To obtain a large contrast between the aperture brightness and the surrounding metal, the metal must be opaque (optically thick), which implies that the film thickness must be several times the skin-depth of the metal. The skin-depth is the penetration depth in the metal at which incoming light intensity has been reduced by $1/e$. Typical skin-depths are of the order of 20 nm for noble metals in the visible spectrum, so film thicknesses of the order of 200 nm are appropriate at optical wavelengths. Even in such thick films, the surface modes on either side of a metal film can also interact via the holes, split and give rise to modes with new energies. Such an effect is especially visible in hole arrays and disappears as the film thickness increases and the modes on either side are decoupled[33,37].

There are many ways to fabricate aperture structures, depending on the scale of interest. For the optical regime, the techniques of choice are: focused ion beam, electron beam lithography and photolithography. The latter two involve several steps but are particularly useful for large-scale structures. The focused ion beam technique, in which the sample is milled by focused ion bombardment, is ideally suited for texturing the metal surface, for instance when preparing grooves around an aperture. Finally, care is needed in preparing the metal films because their quality is an important parameter in the optical properties of the structure.

Potential applications

Surface-wave-activated holes in metal films are finding applications well beyond the illustrations given in the above paragraphs. In the field of opto-electronics for instance, studies are being carried out to extract more light from light-emitting devices[100]. The metal electrodes of such devices, which are normally a source of loss, can be structured with holes to help extract the light from the diode. The need for ever-smaller features on electronic chips is pushing photolithography to use shorter wavelengths, with the associated increased costs and complications. The use of extraordinary optical transmission could perhaps circumvent this problem by using SP-activated lithography masks, which allow subwavelength features in the near-field and high throughput[101–103].

The combination of molecules and holes is another promising area of application, whether for the realization of devices or for the spectroscopic purposes illustrated above. The high optical contrast of SP-activated holes, their small sizes and their simplicity make them ideal candidates for integration on biochips as sensing elements. As in all SP-enhanced phenomena, both the input and output optical fields can be strengthened, with the additional feature that the structure can potentially focus the signal towards a detector. For the purpose of making SP-active devices, the transmission of hole arrays can be switched by controlling the refractive index of molecular materials either electrically[104] or optically up to terahertz speeds[105].

Finally, subwavelength holes might find use in quantum and atom optics. For instance, hole arrays are promising tools in the study of the physical nature—quantum versus classical—of SPs as collective excitations when implemented in quantum entanglement experiments[106,107]. It has been shown theoretically that the extraordinary transmission phenomenon can also be expected for matter waves involving ultracold atoms[108] such as those used in Bose–Einstein condensates. This presents opportunities to create optical elements to manipulate atoms and control their direction.

The potential of the optics of tiny holes in metal screens lies in the contrast between the strong opacity of the metal and the aperture, combined with the fact that the metal allows for high local field enhancements. In addition, the properties of these apertures can be tailored by structuring the metal with modern nanofabrication techniques. The simplicity of the structures and their ease of use should further expand their application in a variety of areas and lead to unsuspected developments.

1. Grimaldi, F.-M. in *Physico-mathesis de Lumine, Coloribus, et Iride, Aliisque Sequenti Pagina Indicatis* 9 (Bologna, 1665).
2. Bethe, H. A. Theory of diffraction by small holes. *Phys. Rev.* **66,** 163–182 (1944).
3. Betzig, E. & Trautman, J. K. Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. *Science* **257,** 189–194 (1992).
4. Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T. & Wolff, P. A. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391,** 667–669 (1998).
5. Ritchie, R. H. Plasma losses by fast electrons in thin films. *Phys. Rev.* **106,** 874–881 (1957).
6. Barnes, W. L., Dereux, A. & Ebbesen, T. W. Surface plasmon subwavelength optics. *Nature* **424,** 824–830 (2003).
7. Roberts, A. Electromagnetic theory of diffraction by a circular aperture in a thick, perfectly conducting screen. *J. Opt. Soc. Am. A* **4,** 1970–1983 (1987).
8. Gordon, R. & Brolo, A. Increased cut-off wavelength for a subwavelength hole in a real metal. *Opt. Express* **13,** 1933–1938 (2005).
9. Obermüller, C. & Karrai, K. Far-field characterization of diffracting apertures. *Appl. Phys. Lett.* **67,** 3408–3410 (1995).
10. Degiron, A., Lezec, H. J., Yamamoto, N. & Ebbesen, T. W. Optical transmission properties of a single subwavelength aperture in a real metal. *Opt. Commun.* **239,** 61–66 (2004).
11. Yin, L. *et al.* Surface palsmons at single nanoholes in Au films. *Appl. Phys. Lett.* **85,** 467–469 (2004).
12. Garcia-Vidal, F. J., Moreno, E., Porto, J. A. & Martin-Moreno, L. Transmission of light through a single rectangular hole. *Phys. Rev. Lett.* **95,** 103901 (2005).
13. Chang, C.-W., Sarychev, A. K. & Shalaev, V. M. Light diffraction by a subwavelength circular aperture. *Laser Phys. Lett.* **2,** 351–355 (2005).
14. Popov, E. *et al.* Surface plasmon excitation on a single subwavelength hole in a metallic sheet. *Appl. Opt.* **44,** 2332–2337 (2005).
15. Webb, K. J. & Li, J. Analysis of transmission through small apertures in conducting films. *Phys. Rev. B* **73,** 033401 (2006).
16. Garcia de Abajo, F. J. Light transmission through a single cylindrical hole in a metallic film. *Opt. Express* **10,** 1475–1484 (2002).
17. Magde, D., Elson, E. & Webb, W. W. Thermodynamic fluctuations in a reacting system - measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* **29,** 705–707 (1972).
18. Levene, M. J. *et al.* Zero-mode waveguides for single molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).
19. Rignault, H. *et al.* Enhancement of single-molecule fluorescence detection in subwavelength apertures. *Phys. Rev. Lett.* **95,** 117401 (2005).

20. Lezec, H. J. *et al.* Beaming light from a subwavelength aperture. *Science* **297,** 820–822 (2002).
21. Thio, T., Pellerin, K. M., Linke, R. A., Lezec, H. J. & Ebbesen, T. W. Enhanced light transmission through a single subwavelength aperture. *Opt. Lett.* **26,** 1972–1974 (2001).
22. Degiron, A. & Ebbesen, T. W. Analysis of the transmission process through single apertures surrounded by periodic corrugations. *Opt. Express* **12,** 3694–3700 (2004).
23. Martin-Moreno, L., Garcia-Vidal, F. J., Lezec, H. J., Degiron, A. & Ebbesen, T. W. Theory of highly directional emission from a single subwavelength aperture surrounded by surface corrugations. *Phys. Rev. Lett.* **90,** 167401 (2003).
24. Garcia-Vidal, F. J., Lezec, H. J., Ebbesen, T. W. & Martin-Moreno, L. Multiple paths to enhance optical transmission through a subwavelength slit. *Phys. Rev. Lett.* **90,** 213901 (2003).
25. Garcia-Vidal, F. J., Martin-Moreno, L., Lezec, H. J. & Ebbesen, T. W. Focusing light with a single subwavelength aperture flanked by surface corrugations. *Appl. Phys. Lett.* **83,** 4500–4502 (2003).
26. Yu, L.-B. *et al.* Physical origin of directional beaming from a subwavelength slit. *Phys. Rev. B* **71,** 041405(R) (2005).
27. Ishi, T., Fujikata, J. & Ohashi, K. Large optical transmission through a single subwavelength hole associated with a sharp-apex grating. *Jpn J. Appl. Phys.* **44,** L170–L172 (2005).
28. Sun, Z. & Kim, H. K. Refractive transmission of light and beam shaping with metallic nano-optics lenses. *Appl. Phys. Lett.* **85,** 642–644 (2004).
29. Gbur, G., Schouten, H. F. & Visser, T. D. Achieving superresolution in near-field optical data readout systems using surface plasmons. *Appl. Phys. Lett.* **87,** 191109 (2005).
30. Fujikata, J. *et al.* Surface plasmon enhancement effect and its application to near-field optical recording. *Trans. Magn. Soc. Jpn* **4,** 255–259 (2004).
31. Nahata, A., Linke, R. A., Ishi, T. & Ohashi, K. Enhanced nonlinear optical conversion from a periodically nanostructured metal film. *Opt. Lett.* **28,** 423–425 (2003).
32. Ishi, T., Fujikata, J., Makita, K., Baba, T. & Ohashi, K. Si nano-photodiode with a surface plasmon antenna. *Jpn J. Appl. Phys.* **44,** L364–L366 (2005).
33. Degiron, A., Lezec, H. J., Barnes, W. L. & Ebbesen, T. W. Effects of hole depth on enhanced light transmission through subwavelength hole arrays. *Appl. Phys. Lett.* **81,** 4327–4329 (2002).
34. Bravo-abad, J. *et al.* How light emerges from an illuminated array of subwavelength holes. *Nature Phys.* **2,** 120–123 (2006).
35. Porto, J. A., Garcia-Vidal, F. J. & Pendry, J. B. Transmission resonances on metalllic gratings with very narrow slits. *Phys. Rev. Lett.* **83,** 2845–2848 (1999).
36. Strelniker, Y. M. & Bergman, D. J. Optical transmission through metal films with a subwavelength hole array in the presence of a magnetic field. *Phys. Rev. B* **59,** R12763 (1999).
37. Martin-Moreno, L. *et al.* Theory of extraordinary optical transmission through subwavelength hole arrays. *Phys. Rev. Lett.* **86,** 1114–1117 (2001).
38. Popov, E., Neviere, M., Enoch, S. & Reinisch, R. Theory of light transmission through subwavelength periodic hole arrays. *Phys. Rev. B* **62,** 16100 (2000).
39. Barbara, A., Quemerais, P., Bustarret, E. & Lopez-Rios, T. Optical transmission through subwavelength metallic gratings. *Phys. Rev. B* **66,** 161403(R) (2002).
40. Baida, F. I. & Van Labeke, D. Light transmission by subwavelength annular aperture arrays in metallic films. *Opt. Commun.* **209,** 17–22 (2002).
41. Sarychev, A. K., Podolskiy, V. A., Dykhne, A. M. & Shalaev, V. M. Resonance transmittance through a metal film with subwavelength holes. *IEEE J. Quant. Elect.* **38,** 956–963 (2002).
42. Sarrazin, M., Vigneron, J. P. & Vigoureux, J.-M. Role of Wood anomalies in optical properties of thin metallic films with a bidimensional array of subwavelength holes. *Phys. Rev. B* **67,** 085415 (2003).
43. Genet, C., van Exter, M. P. & Woerdman, J. P. Fano-type interpretation of red shifts and red tails in hole array transmission spectra. *Opt. Commun.* **225,** 331–336 (2003).
44. Zayats, A. V., Salomon, L. & de Fornel, F. How light gets through periodically nanostructured metal films: a role of surface polaritonic crystals. *J. Microsc.* **210,** 344–349 (2003).
45. Lalanne, P., Rodier, J. C. & Hugonin, J. P. Surface plasmons of metallic surfaces perforated by nanohole arrays. *J. Opt. Pure Appl. Opt.* **7,** 422–426 (2005).
46. Lomakin, V. & Michielssen, E. Enhanced transmission through metallic plates perforated by arrays of subwavelength holes and sandwiched between dielectric slabs. *Phys. Rev. B* **71,** 235117 (2005).
47. Müller, R., Malyarchuk, V. & Lienau, C. Three-dimensional theory on light-induced near-field dynamics in a metal film with a periodic array of nanoholes. *Phys. Rev. B* **68,** 205415 (2003).
48. Takakura, Y. Optical resonance in a narrow slit in a thick metallic screen. *Phys. Rev. Lett.* **86,** 5601–5603 (2001).
49. Shipman, S. P. & Venakides, S. Resonant transmission near nonrobust periodic slab modes. *Phys. Rev. E* **71,** 026611 (2005).
50. Shen, J. T., Catrysse, P. B. & Fan, S. Mechanism for designing metallic metamaterials with a high index of refraction. *Phys. Rev. Lett.* **94,** 197401 (2005).
51. Xie, Y., Zakharian, A. R., Moloney, J. V. & Mansuripur, M. Transmission of light through slit apertures in metallic films. *Opt. Express* **12,** 6106–6121 (2004).
52. Lee, K. G. & Park, Q.-H. Coupling of surface plasmon polaritons and light in metallic nanoslits. *Phys. Rev. Lett.* **95,** 103902 (2005).
53. Marquier, F., Greffet, J.-J., Collin, S., Pardo, F. & Pelouard, J. L. Resonant transmission through metallic film due to coupled modes. *Opt. Express* **13,** 70–76 (2005).
54. Skigin, D. C. & Depine, R. A. Transmission resonances of metallic compound gratings with subwavelength slits. *Phys. Rev. Lett.* **95,** 217402 (2005).
55. Kim, K. Y., Cho, Y. K., Tae, H. S. & Lee, J.-H. Light transmission along dispersive plasmonic gap and its subwavelength guidance characteristics. *Opt. Express* **14,** 320–330 (2006).
56. Liu, W.-C. & Tsai, D. P. Optical tunnelling effect of surface plasmon polaritons and localized surface plasmon resonance. *Phys. Rev. B* **65,** 155423 (2005).
57. Garcia de Abajo, F. J., Saenz, J. J., Campillo, I. & Dolado, J. S. Site and lattice resonances in metallic hole arrays. *Opt. Express* **14,** 7–18 (2006).
58. Chang, S.-H., Gray, S. K. & Schatz, G. C. Surface plasmon generation and light transmission by isolated nanoholes and arrays of nanoholes in thin metal films. *Opt. Express* **13,** 3150–3165 (2005).
59. Bravo-Abad, J., Garcia-Vidal, F. J. & Martin-Moreno, L. Resonant transmission of light through finite chains of subwavelength holes in a metallic film. *Phys. Rev. Lett.* **93,** 227401 (2005).
60. Ghaemi, H. F., Thio, T., Grupp, D. E., Ebbesen, T. W. & Lezec, H. J. Surface plasmons enhance optical transmission through subwavelength holes. *Phys. Rev. B* **58,** 6779–6782 (1998).
61. Sun, Z., Jung, Y. S. & Kim, H. K. Role of surface plasmons in the optical interaction in metallic gratings with narrow slits. *Appl. Phys. Lett.* **83,** 3021–3023 (2003).
62. Barnes, W. L., Murray, W. A., Dintinger, J., Devaux, E. & Ebbesen, T. W. Surface plasmon polaritons and their role in the enhanced transmission of light through periodic arrays of sub-wavelength holes in a metal film. *Phys. Rev. Lett.* **92,** 107401 (2004).
63. Prikulis, J., Hanarp, P., Olofsson, L., Sutherland, D. & Kall, M. Optical spectroscopy of nanometric holes in thin gold films. *Nano Lett.* **4,** 1003–1007 (2004).
64. Klein Koerkamp, K. J., Enoch, S., Segerink, F. B., van Hulst, N. F. & Kuipers, L. Strong influence of hole shape on extraordinary transmission through periodic arrays of subwavelength holes. *Phys. Rev. Lett.* **92,** 183901 (2004).
65. Degiron, A. & Ebbesen, T. W. The role of localized surface plasmon modes in the enhanced transmission of periodic subwavelength apertures. *J. Opt. Pure Appl. Opt.* **7,** S90–S96 (2005).
66. Gordon, R. *et al.* Strong polarization in the optical transmission through elliptical nanohole arrays. *Phys. Rev. Lett.* **92,** 037401 (2004).
67. Ye, Y.-H. & Zhang, J.-Y. Enhanced light transmission through cascaded metal films perforated with periodic hole arrays. *Opt. Lett.* **30,** 1521–1523 (2005).
68. Krasavin, A. V. *et al.* Polarization conversion and "focusing" of light propagating through a small chiral hole in a metallic screen. *Appl. Phys. Lett.* **86,** 201105 (2005).
69. Wang, Q.-J., Li, J.-Q., Huang, C.-P., Zhang, C. & Zhu, Y.-Y. Enhanced optical transmission through metal films with rotation-symmetrical hole arrays. *Appl. Phys. Lett.* **87,** 091105 (2005).
70. Ropers, C. *et al.* Femtosecond light transmission and subradiant damping in plasmonic crystals. *Phys. Rev. Lett.* **94,** 113901 (2005).
71. Dogariu, A., Thio, T., Wang, L. J., Ebbesen, T. W. & Lezec, H. J. Delay in light transmission through small apertures. *Opt. Lett.* **26,** 450–452 (2001).
72. Halté, V., Benabbas, A., Guidoni, L. & Bigot, J.-Y. Femtosecond dynamics of the transmission of gold arrays of subwavelength holes. *Phys. Status Solidi (b)* **242,** 1872–1876 (2005).
73. Dechant, A. & Elezzabi, A. Y. Femtosecond optical pulse propagation in subwavelength metallic slits. *Appl. Phys. Lett.* **84,** 4678–4680 (2004).
74. Kwak, E.-S. *et al.* Surface plasmon standing waves in large-area subwavelength hole arrays. *Nano Lett.* **5,** 1963–1967 (2005).
75. Kim, D. S. *et al.* Microscopic origin of surface-plasmon radiation in plasmonic band-gap nanostructures. *Phys. Rev. Lett.* **91,** 143901 (2003).
76. Egorov, D., Dennis, B. S., Blumberg, G. & Haftel, M. I. Two-dimensional control of surface plasmons and directional beaming from arrays of subwavelength apertures. *Phys. Rev. B* **70,** 033404 (2004).
77. Chyan, J. Y., Chang, C. A. & Yeh, J. A. Development and characterization of a broad-bandwidth polarization-insensitive subwavelength optical device. *Nanotechnology* **17,** 40–44 (2006).
78. Schouten, H. F. *et al.* Plasmon-assisted two-slit transmission: Young's experiment revisited. *Phys. Rev. Lett.* **94,** 053901 (2005).
79. Altewischer, E., van Exter, M. P. & Woerdman, J. P. Polarization analysis of propagating surface plasmons in a subwavelength hole array. *J. Opt. Soc. Am. B* **20,** 1927–1931 (2003).
80. Williams, S. M. *et al.* Use of the extraordinary infrared transmission of metallic subwavelength arrays to study the catalyzed reaction of methanol to formaldehyde on copper oxide. *J. Phys. Chem. B* **108,** 11833–11837 (2004).
81. Brolo, A. G. *et al.* Enhanced fluorescence from arrays of nanoholes in a gold film. *J. Am. Chem. Soc.* **127,** 14936–14941 (2005).
82. Liu, Y., Bishop, J., Williams, L., Blair, S. & Herron, J. Biosensing based upon molecular confinement in a metallic nanocavity arrays. *Nanotechnology* **15,** 1368–1374 (2004).
83. Brolo, A. G., Gordon, R., Leathem, B. & Kavanagh, K. L. Surface plasmon sensor based on the enhanced light transmission through arrays of nanoholes in gold films. *Langmuir* **20,** 4813–4815 (2004).
84. Moran, C. E., Steele, J. M. & Halas, N. J. Chemical and dielectric manipulation of plasmonic band gap of metallodielectric arrays. *Nano Lett.* **4,** 1497–1500 (2004).

85. Stark, P. R. H., Halleck, A. E. & Larson, D. N. Short order nanohole arrays in metals for highly sensitive probing of local indices of refraction as the basis for a highly multiplexed biosensor technology. *Methods* **37,** 37–47 (2005).
86. Brolo, A. G., Arctander, E., Gordon, R., Leathem, B. & Kavanagh, K. L. Nanohole-enhanced Raman scattering. *Nano Lett.* **4,** 2015–2018 (2004).
87. Williams, S. M. *et al.* Scaffolding for nanotechnology: extraordinary infrared transmission of microarrays for stacked sensors and surface spectroscopy. *Nanotechnology* **15,** S495–S503 (2004).
88. Coe, J. V. *et al.* Extra IR transmission with metallic arrays of subwavelength holes. *Anal. Chem.* **78,** 1385–1389 (2006).
89. Rindzevicius, T. *et al.* Plamsonic sensing characteristics of single nanometric holes. *Nano Lett.* **5,** 2335–2339 (2005).
90. Dintinger, J., Klein, S. & Ebbesen, T. W. Molecule–surface plasmon interactions in hole arrays: enhanced absorption, refractive index changes and all-optical switching. *Adv. Mat.* **18,** 1267–1270 (2006).
91. Gomez Rivas, J., Schotsch, C., Haring Bolivar, P. & Kurz, H. Enhanced transmission of Thz radiation through subwavelength holes. *Phys. Rev. B* **68,** 201306(R) (2003).
92. Shou, X., Agrawal, A. & Nahata, A. Role of metal thickness on the enhanced transmission properties of a periodic array of subwavelength apertures. *Opt. Express* **13,** 9834–9840 (2005).
93. Lockyear, M. J., Hibbins, A. P. & Sambles, J. R. Surface-topography-induced enhanced transmission and directivity of microwave radiation through a subwavelength circular metal aperture. *Appl. Phys. Lett.* **84,** 2040–2042 (2004).
94. Pendry, J. B., Martin-Moreno, L. & Garcia-Vidal, F. J. Mimicking surface plasmons with structured surfaces. *Science* **305,** 847–848 (2004).
95. Garcia-Vidal, F. J., Martin-Moreno, L. & Pendry, J. B. Surfaces with holes in them: new plasmonic metamaterials. *J. Opt. Pure Appl. Opt.* **7,** S97–S101 (2005).
96. Lalanne, P. & Hugonin, J. P. Interaction between optical nano-objects at metallo-dielectric interfaces. *Nature Phys.* **2,** 551–556 (2006).
97. Visser, T. D. Surface plasmons at work? *Nature Phys.* **2,** 509–510 (2006).
98. Gruhlke, R., Hod, W. & Hall, D. Surface-plasmon cross coupling in molecular fluorescence near a corrugated thin film. *Phys. Rev. Lett.* **56,** 2838–2841 (1986).
99. Bonod, N., Enoch, S., Li, L., Popov, E. & Nevière, M. Resonant optical transmission through thin metallic films with and without holes. *Opt. Express* **11,** 482–490 (2003).
100. Liu, C., Kamaev, V. & Vardeny, Z. V. Efficency enhancement of an organic light-emitting diode with a cathode forming two-dimensional periodic hole array. *Appl. Phys. Lett.* **86,** 143501 (2005).
101. Srituravanich, W., Fang, N., Sun, C., Luo, Q. & Zhang, X. Plasmonic nanolithography. *Nano Lett.* **4,** 1085–1088 (2004).
102. Luo, X. & Ishihara, T. Sub-100nm photolithography based on plasmon resonance. *Jpn J. Appl. Phys.* **43,** 4017–4021 (2004).
103. Shao, D. B. & Che, S. C. Surface-plasmon-assisted nanoscale photolithography by polarized light. *Appl. Phys. Lett.* **86,** 253107 (2005).
104. Kim, T. J., Thio, T., Ebbesen, T. W., Grupp, D. E. & Lezec, H. J. Control of optical transmission through metals perforated with subwavelength hole arrays. *Opt. Lett.* **24,** 256–258 (1999).
105. Dintinger, J., Robel, I., Kamat, P. V., Genet, C. & Ebbesen, T. W. Terahertz all-optical molecule-plasmon modulation. *Adv. Mater.* **18,** 1645–1648 (2006).
106. Altewisher, E., van Exter, M. P. & Woerdman, J. P. Plasmon-assisted transmission of entangled photons. *Nature* **418,** 304–306 (2002).
107. Fasel, S. *et al.* Energy-time entanglement preservation in plasmon-assisted light transmission. *Phys. Rev. Lett.* **94,** 110501 (2005).
108. Moreno, E., Fernandez-Dominguez, A. I., Cirac, I. J., Garcia-Vidal, F. J. & Martin-Moreno, L. Resonant transmission of cold atoms through subwavelength apertures. *Phys. Rev. Lett.* **95,** 170406 (2005).

Acknowledgements Our research was supported by the European Community, Network of Excellence PLASMO-NANO-DEVICES, STREP SPP, the ANR grant COEXUS, the CNRS, and the French Ministry of Higher Education and Research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to T.W.E. (ebbesen@isis-ulp.org).

ARTICLES

Oxygen content of transmembrane proteins over macroevolutionary time scales

Claudia Acquisti[1]†, Jürgen Kleffe[2] & Sinéad Collins[1]

We observe that the time of appearance of cellular compartmentalization correlates with atmospheric oxygen concentration. To explore this correlation, we predict and characterize the topology of all transmembrane proteins in 19 taxa and correlate differences in topology with historical atmospheric oxygen concentrations. Here we show that transmembrane proteins, individually and as a group, were probably selectively excluding oxygen in ancient ancestral taxa, and that this constraint decreased over time when atmospheric oxygen levels rose. As this constraint decreased, the size and number of communication-related transmembrane proteins increased. We suggest the hypothesis that atmospheric oxygen concentrations affected the timing of the evolution of cellular compartmentalization by constraining the size of domains necessary for communication across membranes.

One of the major transitions in macroevolution was the appearance of eukaryotic cells between 2.1 and 1.8 billion years ago[1–3]. Cellular compartmentalization by membranes that are impermeable to large or charged molecules requires transport and communication across intracellular membranes. Eukaryotes devote more proteins to roles in communication than prokaryotes; this innovation involved a shift in the dominant secondary structures of transmembrane proteins[4]. Protein secondary structure is largely determined by hydrophobicity[5], where oxygen and nitrogen are vital to forming hydrophilic residues. Transmembrane protein topology is further influenced by charge, where positively charged amino acids are more prevalent in cytoplasmic domains and negatively charged amino acids are more prevalent in extracellular domains[6–8]. This implies that changes in protein atomic composition may occur in parallel with changes in protein function. Traditionally, functional changes were thought to be associated with changes in amino acid sequence[9], but an alternative approach is to consider proteins at the atomic level. This may be appropriate when large fluctuations in the elemental components of proteins occur through changes in absolute abundance, relative abundance, or form. In this case, nutritional constraints, metabolic optimization and chemical properties such as redox state may have important roles in protein evolution.

The atomic content of biomolecules has a role in evolution

Several examples of stoichiometric constraints on evolutionary and ecological outcomes have been reported recently. For example, variation in the atomic content of proteins in cyanobacterial light-harvesting proteins and microbial sulphur assimilatory enzymes correlates with nutrient availability[10,11]. Similarly, the carbon content of proteomes differs between species and correlates with genomic G+C content, which may reflect carbon availability in natural habitats[12]. The nitrogen content of proteins is lower in plants than in animals and is related to gene expression levels in plants[13]. These studies indicate that physiology, proteomes and genomes may bear detectable ecological imprints over macroevolutionary time scales, and that ancient habitat composition may affect current proteome composition. Furthermore, recent advances in the field of ecological stoichiometry have shown that the relative abundances of phosphorus to carbon or nitrogen can influence ecological outcomes. This may occur through stoichiometric constraints on growth rate, which can lead to variation in life-history traits that subsequently affect species–species interactions[14,15]. The relative abundances of nutrients have been associated with many macroevolutionary innovations including the appearance of winged insects[16] and the timing of the Cambrian explosion[17].

Molecular oxygen was introduced relatively quickly into the atmosphere by the ancestors of cyanobacteria about 2.2 billion years ago[18] and has varied between about 15 and 35% over the time that eukaryotic cells have been present[19]. Here we address how atmospheric oxygen concentrations may have constrained transmembrane protein composition and structure, and then suggest a functional interpretation of these constraints at a cellular level. We explore the possible macroevolutionary consequences of this—namely, the timing of the appearance of eukaryotic cells. First, we characterize the oxygen content and topology of the entire set of predicted transmembrane proteins from 19 organisms (listed in Supplementary Table 1). We then show how the oxygen content of transmembrane proteins varies with respect to atmospheric oxygen concentration over the past 3.5 billion years, and we suggest a mechanism of how this could have constrained the timing of evolution of cellular compartmentalization.

Oxygen content of transmembrane proteins in prokaryotes and eukaryotes

To investigate how atomic content changed over macroevolutionary time scales, we calculated the mean side-chain density for carbon, hydrogen, oxygen and nitrogen content for the full, predicted proteomes of 19 organisms. Carbon and hydrogen content density functions have a nearly gaussian distribution. In contrast, we found that

[1]Max Planck Institute for Plant Breeding Research, Carl-von-Linnè-Weg 10, 50829 Köln, Germany. [2]Institute of Molecular Biology and Biochemistry, Charité Campus Benjamin Franklin, Arnimallee 22, 14195 Berlin, Germany. †Present address: Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301, USA.

the density function for oxygen content is bimodal in prokaryotes and nearly gaussian in eukaryotes, with the exception of *Caenorhabditis elegans* and *Giardia lamblia* (Fig. 1; Supplementary Figs 1, 2). The same pattern occurs for nitrogen content density, although it is less pronounced (Supplementary Figs 3–5). To test if the low-oxygen peak of the bimodal distributions was associated with a particular subset of proteins, we extracted the transmembrane proteins from the full proteome for each organism using a transmembrane protein topology prediction method that was based on a hidden Markov model (TMHMM)[20]. We assigned the remainder of the full proteome minus the putative transmembrane proteins to the 'non-transmembrane protein' group. Comparing the oxygen content density functions for the non-transmembrane with those for the transmembrane protein set shows that the distributions of the two sets are significantly distinct from each other (chi-squared test, $P < 0.0001$; Fig. 1).

Transmembrane proteins constitute a separate group in terms of oxygen content density in eukaryotic and prokaryotic proteomes, although the difference is more pronounced in prokaryotes. The difference in mean oxygen percentages of transmembrane and non-transmembrane proteins is 0.73 in eukaryotes, whereas it is 1.33 in prokaryotes. Transmembrane proteins have lower oxygen contents than non-transmembrane proteins in each taxonomic domain (Cochran-Cox 1-tailed *t*-test, $P < 0.0005$). Although eukaryotic proteomes are larger than prokaryotic ones, the fraction of transmembrane proteins is largely conserved: in most genomes, 20–30% of all genes encode transmembrane proteins[21]. This suggests that the differences in oxygen content density distributions between prokaryotes and eukaryotes are not attributable to gross differences in proteome composition. However, individual transmembrane proteins in eukaryotes tend to be longer than in prokaryotes[22], and often

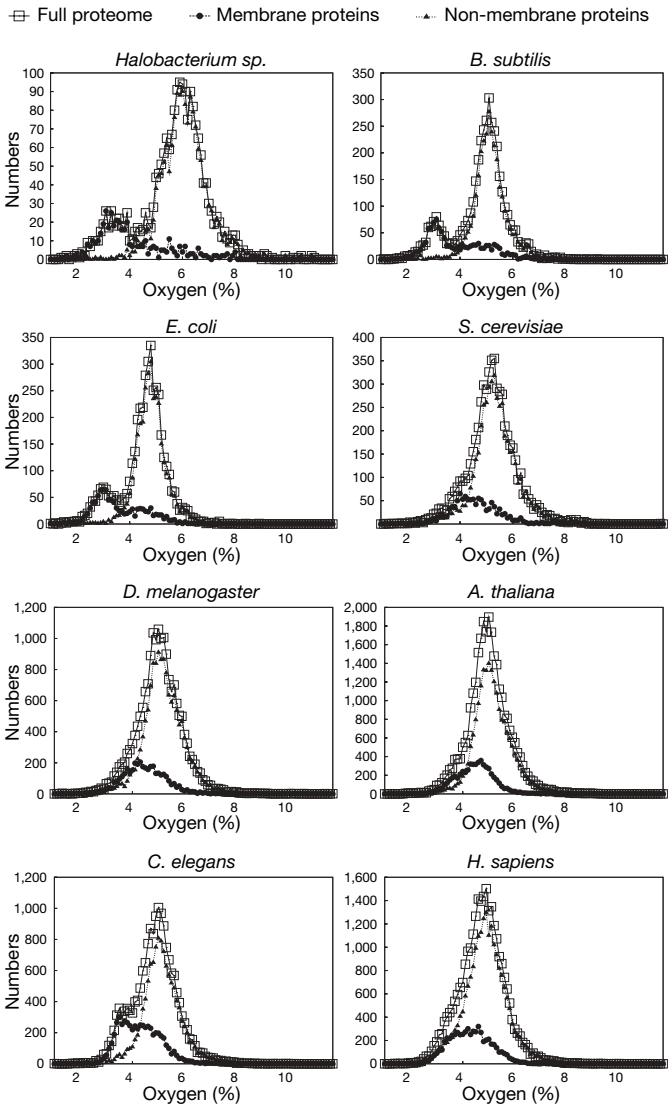


Figure 1 | Oxygen content density functions on full proteome, transmembrane and non-transmembrane protein sets. For each proteome (*Halobacterium* sp., *Bacillus subtilis*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Homo sapiens*) the oxygen content density was calculated as the percentage of oxygen atoms found in residue side chains for each protein, and the percentages plotted as a histogram. The oxygen content density histograms were plotted separately for the transmembrane and non-transmembrane protein sets. Sample sizes given in Supplementary Table 1.

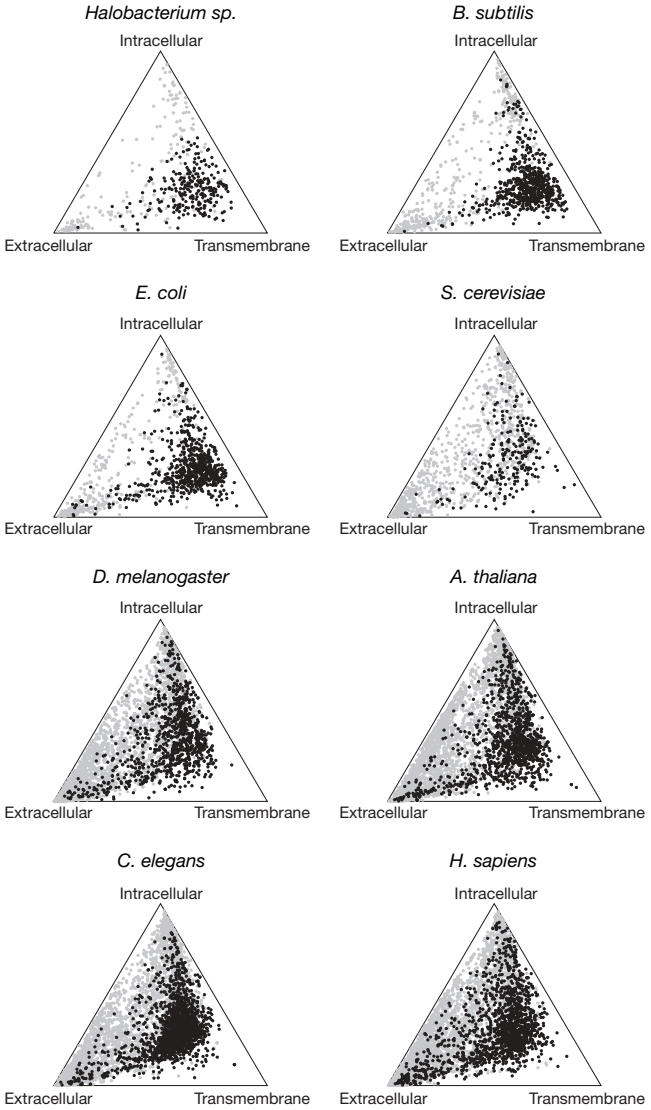


Figure 2 | Ternary diagrams of compositional data for transmembrane, extracellular and intracellular domains for the entire predicted transmembrane protein set. Each individual protein is represented by a three-dimensional vector of components (extracellular, intracellular and transmembrane), the sum of which is 1. The magnitude of a component is equal to the length of the perpendicular axis leading to the edge opposite the vertex having the same identity as that component. The coordinates of each data point show topology, and colour shows oxygen content (dark colour, [O] < 3.9%); light colour, [O] ≥ 3.9%. Sample sizes given in Supplementary Table 1.

lengthen either the intra- or extracellular domains relative to prokaryotes, rather than simply expanding the entire protein[23].

Oxygen content and transmembrane protein topology

To describe the relationship between transmembrane protein oxygen content and their topology, and to further characterize differences between the prokaryotic and eukaryotic transmembrane proteins, the topology of each transmembrane protein was determined for the entire proteome of each organism using TMHMM[20] (Fig. 2; Supplementary Figs 6–7). Two clusters can be seen within each diagram. One is made up of proteins with a high proportion of transmembrane domains that have low oxygen content, and the second is made up of proteins with a lower proportion of transmembrane domains with higher oxygen content. The distributions of the low-oxygen and high-oxygen proteins in the ternary diagrams differ significantly from each other (multivariate likelihood ratio test for compositional data[24], $P = 0.00001$). The overlap between protein sets defined either by oxygen content or by topology is at least 80% in all cases, indicating that there is good agreement between oxygen content and topology when the two are independently estimated. Our distribution of topologies is consistent with previous work showing that transmembrane proteins have a tendency to either form many transmembrane domains with short connecting loops or few transmembrane domains with large extracellular domains, but not both[22]. In addition, our compositional data show differences in transmembrane protein topology between eukaryotes and prokaryotes, with prokaryotes having a higher proportion of transmembrane-domain-rich proteins. Similarly, there is a threefold difference in the number of low-oxygen transmembrane proteins between prokaryotes and eukaryotes, whereas there is a tenfold difference in the number of high-oxygen transmembrane proteins, indicating that high-oxygen proteins dominated by intra- or extracellular domains are preferentially added as proteomes increase in size over time. The preferential addition of high-oxygen proteins over time shows that some transmembrane proteins may have been oxygen-limited at the time they evolved, or that there was selection against the use of oxygen in external transmembrane domains. This indicates that changes in transmembrane protein oxygen content may have been connected with changes in protein structure and function over geological time.

The topology description above is normalized and reflects differences in proportional contributions of inside (intracellular), outside (extracellular) and transmembrane domains to total protein length. However, it does not describe how the absolute sizes of these domains differ systematically between organisms. Within sets of high- and low-oxygen transmembrane proteins, we estimated the mean number of residues devoted to the inside, outside and transmembrane domains per proteome, on the basis of TMHMM (Fig. 3). Transmembrane proteins are either composed of many transmembrane segments with few short loops, or few transmembrane segments with large intra- or extracellular loops[22]. These topologies correspond to channels and receptors, respectively. We also calculated the oxygen concentration of each domain (Fig. 4), finding that high-oxygen transmembrane proteins (receptors) had higher oxygen content in their external than in their internal and transmembrane domains, or compared with any domain of the low-oxygen transmembrane proteins (Cochran-Cox 1-tailed *t*-test, corrected $P = 0.005$). When the domain lengths are plotted against the time of appearance of class,

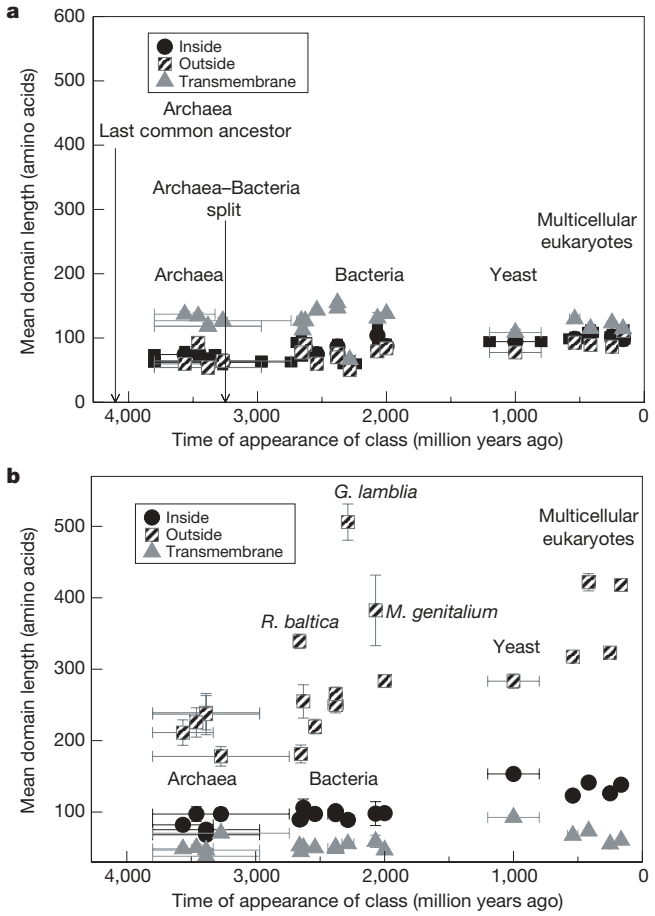


Figure 3 | Mean domain length versus time of appearance of class. **a**, Low-oxygen transmembrane proteins ([O] < 3.9%). Mean values for domain lengths of outside and transmembrane domains do not correlate with time (Spearman's coefficients: inside, $r_s = -0.542$, $P = 0.009$; outside, $r_s = -0.229$, $P = 0.17$; transmembrane, $r_s = 0.217$, $P = 0.19$). The ranges of lengths are: 60<inside<109; 50<outside<111; 66.5<transmembrane<154.7 amino acids. **b**, High-oxygen transmembrane proteins ([O] ≥ 3.9%). Inside, outside and transmembrane domain lengths correlate with time (Spearman's coefficients: inside, $r_s = -0.802$, $P = 0.00002$; outside, $r_s = -0.740$, $P = 0.0001$; transmembrane, $r_s = 0.606$, $P = 0.003$). The ranges of lengths are: 68<inside<150; 177<outside<505; 38<transmembrane<92 amino acids). Points show mean ± s.e.m. of the length, and the range of time of appearance[30–34]. Sample sizes given in Supplementary Table 1.

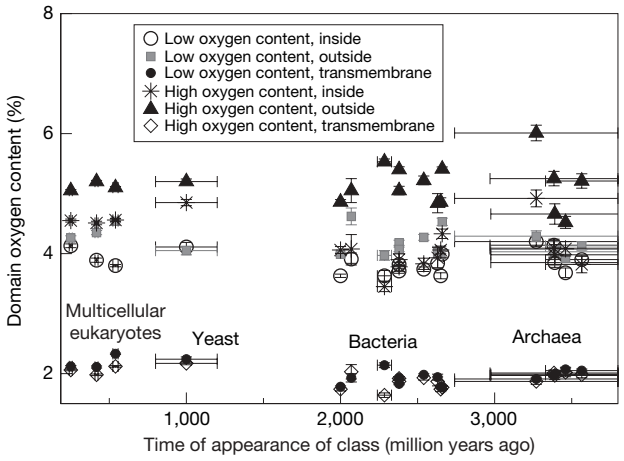


Figure 4 | Inside, outside and transmembrane domain oxygen content versus time of appearance of class. For each transmembrane protein with low-oxygen ([O] < 3.9%) or high-oxygen ([O] ≥ 3.9%) content, the oxygen content of different domains was estimated using TMHMM[20]. Each point represents the mean ± s.e.m. of oxygen content, and the range of time of predicted first appearance[30–34]. The mean oxygen content of specific domains does not correlate with the time of appearance of class (Spearman's coefficients: low-oxygen domains, inside, $r_s = -0.125$, $P = 0.9$; outside, $r_s = -0.22$, $P = 0.37$; transmembrane, $r_s = -0.404$, $p = 0.11$; and high-oxygen domains, inside, $r_s = -0.350$, $P = 0.15$, outside, $r_s = 0.107$, $P = 0.65$, transmembrane, $r_s = -0.071$, $P = 0.77$). Sample sizes given in Supplementary Table 1.

high-oxygen transmembrane proteins show a rapid increase in outside relative to inside domains (Fig. 3b). In contrast, low oxygen transmembrane proteins do not show any obvious difference in rates of increase, and mean values of inside and outside domains change relatively slowly over time in low-oxygen transmembrane proteins (Fig. 3a). However, the mean oxygen concentration of the external domains did not change directionally over time in either the high- or low-oxygen transmembrane proteins (receptors or channels) (Spearman's coefficient: low-oxygen proteins $r_s = -0.22$, $P > 0.37$; high-oxygen proteins $r_s = 0.107$, $P = 0.65$), meaning that the charge per unit length remained relatively constant (Fig. 4). This is consistent with charge density being important for the insertion of domains into cellular membranes[6–8], and suggests that the total oxygen content of a domain must be altered by changing the length of that domain. The relatively rapid changes in the size of the oxygen-rich external domains coincide with increasing organismal complexity, whereas changes in the nitrogen-rich internal domains are not as pronounced (Fig. 3). Differential rates of evolution have previously been observed in the outside and inside domains of chemokine receptors[23], which is consistent with the general trend we found for all high-oxygen transmembrane proteins. Interestingly, the parasitic bacteria *Mycoplasma genitalium* has much longer external domains than expected, which are similar in length to eukaryotes, perhaps as a result of co-evolution with eukaryotic hosts (Fig. 3). The compartmentalized prokaryote *Rhodopirellula baltica* also has longer extracellular domains than the other prokaryotes, indicating that external domains in transmembrane proteins may be important to this compartmentalization (Fig. 3).

In our data, the high-oxygen transmembrane proteins show an increase in receptors relative to channels in eukaryotes, coupled with a faster rate of growth of external than internal domains over millions of years. This is consistent with previous observations that the percentage of total open reading frames encoding transport proteins is lower in eukaryotes than in prokaryotes[4]. This increase is correlated with the transitions from prokaryotic to eukaryotic cells, and from unicellular to multicellular organisms[25,26]. One possible explanation for different domain growth rates is that the domains may have different elemental requirements, with the positively charged internal domains requiring more nitrogen, and the negatively charged external domains requiring more oxygen. There is a similar, yet weaker signal in the overall per residue nitrogen content, which suggests that oxygen levels were more important to the changes in transmembrane protein topology seen here than changes in nitrogen levels. Still, it does not exclude the possibility that the ratio of the two elements had a role.

Taken together, data in Fig. 3 and in Fig. 4 show that the oxygen content of extracellular domains is higher in receptors than it is in channels; this is evident from increases in both the density of oxygen content and the total length of external domains. This points towards a key role for oxygen in the increase in abundance and size of receptors over time. In terms of cellular function, a faster increase in the size of external domains indicates an increase in the proportion of communication-related proteins over time. The differential changes in domain length between high- and low-oxygen transmembrane proteins, together with previous work, suggest that external loops of transmembrane proteins are under different selective constraints than internal loops. This constraint may have prevented the formation of large external domains, limiting communication across membranes when atmospheric oxygen concentrations were low.

Atmospheric oxygen levels, cellular communication and compartmentalization

If atmospheric oxygen levels constrained transmembrane protein composition over very long time scales, it is expected that the two should co-vary. We found that the atomic composition of transmembrane proteins does scale with atmospheric oxygen levels over macroevolutionary time scales. Figure 5a shows the mean oxygen contents of high- and low-oxygen transmembrane proteins separately in 19 organisms plotted against the atmospheric oxygen concentration at the time the classes of these organisms first appeared. Archaea, and, to a lesser extent, other prokaryotes, show a similar mean oxygen content of high-oxygen transmembrane proteins and a lower mean oxygen content of low-oxygen transmembrane proteins relative to eukaryotes (Spearman's coefficients: low-oxygen transmembrane proteins, $r_s = 0.601$, $P = 0.0032$; high-oxygen transmembrane proteins, $r_s = -0.114$, $P = 0.33$). This suggests that oxygen availability influenced transmembrane protein composition, and that older taxa exclude oxygen from low-oxygen proteins to a greater extent than do younger taxa.

The selective use of oxygen by older taxa indicates that atmospheric oxygen concentrations may have limited the size and/or number of high-oxygen proteins that were produced in ancient proteomes. Figure 5b shows the ratio of the numbers of high- and low-oxygen transmembrane proteins. In contrast with the composition of individual transmembrane proteins (Fig. 5a), where single-celled compartmentalized organisms (*Saccharomyces cerevisiae*, *R. baltica* and *G. lamblia)* are intermediate between prokaryotes and multicellular eukaryotes, these organisms show the same properties as other eukaryotes in terms of the proportion of the proteome devoted to high-oxygen transmembrane proteins, regardless of whether they are prokaryotic or eukaryotic (Fig. 5b). In addition, eubacteria and

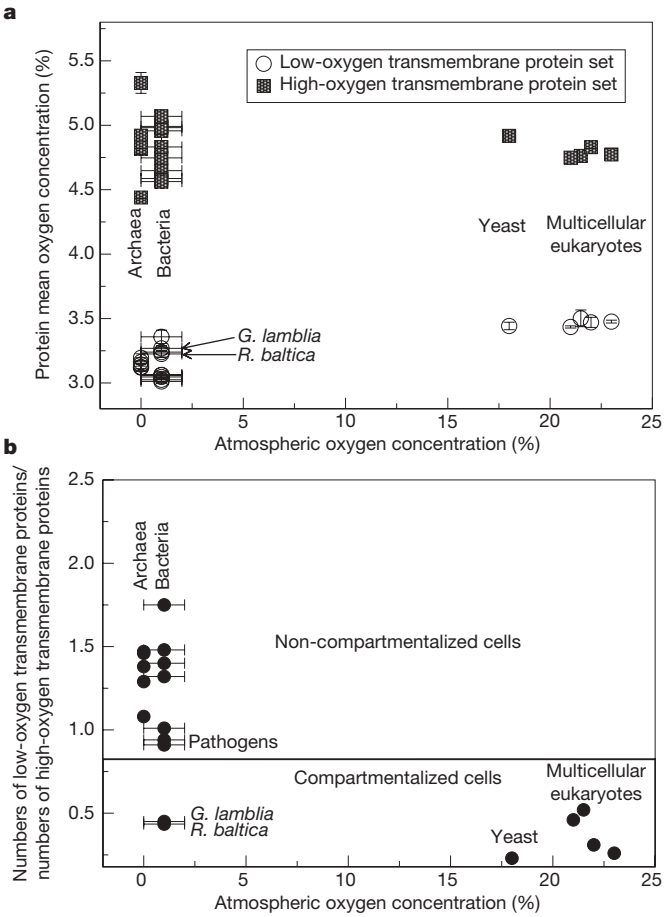


Figure 5 | Mean proteome oxygen content. a, Mean oxygen content of amino acid side chains in high-oxygen ($[O] \geq 3.9\%$) and low-oxygen ($[O] < 3.9\%$) transmembrane proteins plotted against the atmospheric oxygen level at the time of appearance of the organism's class. Each point represents the mean ± s.e.m. of the oxygen content and the range of atmospheric oxygen concentrations at the time of appearance[18,19]. **b**, Ratio of the absolute numbers of low- to high-oxygen transmembrane proteins per proteome versus atmospheric oxygen concentration at the time of appearance. Error bars show the range of atmospheric oxygen concentrations at the time of appearance[18,19]. Sample sizes are given in Supplementary Table 1.

archaea are indistinguishable in this respect, and there is no systematic association between eukaryote complexity, proteome oxygen partitioning, and atmospheric oxygen levels. This suggests a very basic functional difference associated with proteome composition. Because the only two groups that can be distinguished by how oxygen is partitioned at the proteome level are compartmentalized and non-compartmentalized cells ($F_{(1,14)} = 13.25$, $P = 0.002$), the simplest interpretation of this observation is that cellular compartmentalization requires larger intra- and extracellular domains, probably in order to integrate cellular processes such as signalling and transport. In an oxygen-poor (reducing) atmosphere, it may not have been possible to produce a large enough number of these domains for communication in a compartmentalized cell. As such, atmospheric oxygen concentration may have affected the rate of increase in cellular complexity and the timing of the appearance of eukaryotic cells.

Discussion

In this study, we have shown that transmembrane proteins can be divided into two groups according to their oxygen content. Independent topology prediction reveals these same two groups. We have shown that the proportion of receptors to channels increases over time and coincides with a change in cellular organization. In addition, older proteomes contain less oxygen per residue and produce fewer high-oxygen proteins. Taken together, this suggests that oxygen use was selected against in these proteomes. This constraint lessened over time as the concentration of atmospheric oxygen increased, which resulted in the extracellular domains of transmembrane proteins increasing in size over time faster than the internal domains. Consequently, we propose the following hypothetical mechanism: atmospheric oxygen concentration constrained the topology of ancient transmembrane proteins by limiting the number and size of external domains that could be formed.

Any mechanistic explanation of how atmospheric oxygen concentration limited the number and size of external domains is necessarily speculative. One possibility is that it was simply futile to exude large, oxygen-rich domains in a reducing atmosphere where oxidized amino acids could have been rapidly reduced. In this case, the use of oxygen-rich amino acids would have been selected against by natural selection because protein structure would have been more robust when fewer oxidized residues were exuded. Linking this to the timing of appearance of eukaryotic cells implies that the oxygen content is preferentially increased in receptors, and that this increase affects receptor function. This makes intuitive sense because the external domains of receptors required for communication have specific secondary and tertiary structures, many of which have some minimum size[23]. This is consistent with the bias we found towards having both longer and more oxygen-dense external domains in receptors relative to channels, and with the fact that eukaryotic genomes encode more and larger receptors than do prokaryotes. This suggests that protein oxygen content itself is important, rather than being a proxy for some other property. A second possibility is metabolic limitation. There is less evidence for direct limitation, though the synthesis of tyrosine requires molecular oxygen; indirect limitation seems more likely. For example, the synthesis of amino acids with oxygen in their side chains requires less energy than that of other amino acids[27], so selectively excluding oxygen-rich amino acids entails an energetic cost. In addition, the synthesis of many hormones and neurotransmitters requires molecular oxygen, indicating that high levels of molecular oxygen may be needed for communication-related molecules in general.

Constraints on transmembrane protein topology may have played an important part in the timing of the appearance of compartmentalized cells. One of the limits inherent in using data about extant organisms to draw conclusions about constraints on their ancestors is that it is impossible to know how much adaptation to current habitat may bias these data. However, it is reasonable to assume that some historical imprint remains in sequence data. We used organisms with a wide range of metabolism and cellular organization. We have used several different archaea and an extremophile bacterium (*Aquifex aeolicus*), such that the bias associated with any particular extreme of environment is taken into account. Many of the microbes have experienced high levels of oxygen relative to their ancestors for several million years, such that the differences we report are a conservative estimate of the extent to which oxygen may have historically limited proteome composition. This work adds to a growing body of literature connecting atmospheric oxygen levels with macroevolutionary changes, most recently with complexity in metabolic networks[28] and cell types[29]. In order to understand the broad role of oxygen levels in major transitions, further investigation and cooperation between the fields of palaeoclimatology, evolutionary biology and bioinformatics are necessary. This promises to yield many interesting results that address fundamental relationships between macroevolutionary transitions and environmental change.

Received 8 June; accepted 14 November 2006.
Published online 20 December 2006.

1. Han, T. M. & Runnegar, B. Macroscopic eukaryotic algae from the 2.1-billion-year-old Negaunee iron-formation, Michigan. *Science* **257,** 232–235 (1992).
2. Schopf, J. W. & Klein, C. (eds) *The Proterozoic Biosphere: a Multidisciplinary Study* (Cambridge Univ. Press, New York, 1992).
3. Knoll, A. H. Proterozoic and early Cambrian protists: evidence for accelerating evolutionary tempo. *Proc. Natl Acad. Sci. USA.* **91,** 6743–6750 (1994).
4. Ren, Q. & Paulsen, Q. T. Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comp. Biol.* **1,** e27 (2005).
5. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
6. von Heijne, G. Net N–C charge imbalance may be important for signal sequence function in bacteria. *J. Mol. Biol.* **192,** 287–290 (1986).
7. Sipos, L. & von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **213,** 1333–1340 (1993).
8. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rules. *J. Mol. Biol.* **225,** 487–494 (1992).
9. Tourasse, N. J. & Li, W. H. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* **17,** 656–664 (2000).
10. Mazel, D. & Marlière, P. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature* **341,** 245–248 (1989).
11. Baudouin-Cornu, P., Surdin-Kerjan, Y., Marlière, P. & Thomas, D. Molecular evolution of protein atomic composition. *Science* **293,** 297–300 (2001).
12. Baudouin-Cornu, P., Schuerer, K., Marlière, P. & Thomas, D. Intimate evolution of Proteins. Proteome atomic content correlates with genome base composition. *J. Biol. Chem.* **279,** 5421–5428 (2004).
13. Elser, J. J., Fagan, W. F., Subramanian, S. & Kumar, S. Signatures of ecological resource availability in the animal and plant proteomes. *Mol. Biol. Evol.* **23,** 1946–1951 (2006).
14. Elser, J. J. *et al.* Biological stoichiometry from genes to ecosystems. *Ecol. Lett.* **3,** 540–550 (2000).
15. Kay, A. D. *et al.* Toward a stoichiometric framework for evolutionary biology. *Oikos* **109,** 6–17 (2005).
16. Hasenfuss, I. A possible evolutionary pathway to insect flight starting from lepismatid organization. *J. Zool. Syst. Evol. Res.* **40,** 65–81 (2002).
17. Elser, J. J., Watts, J., Schampel, J. H. & Farmer, J. Early Cambrian food webs on a trophic knife-edge? A hypothesis and preliminary data from a modern stromatolite-based ecosystem. *Ecol. Lett.* **9,** 295–303 (2006).
18. Knoll, A. H. *Life on a Young Planet: The First Three Billion Years of Evolution on Earth* (Princeton Univ. Press, Princeton and Oxford, 2003).
19. Berner, R. A. Atmospheric oxygen over Phanerozoic time. *Proc. Natl Acad. Sci. USA.* **96,** 10955–10957 (1999).
20. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
21. Stevens, T. J. & Arkin, I. T. Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins* **39,** 417–420 (2000).
22. Wallin, E. & von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7,** 1029–1038 (1998).
23. Liò, P. & Vannucci, M. Investigating the evolution and structure of chemokine receptors. *Gene* **317,** 29–37 (2003).
24. Aitchison, J. *The Statistical Analysis of Compositional Data* 154–155 (Chapman and Hall, London, 1986).
25. Tamames, J., Ouzounis, C., Sander, C. & Valencia, A. Genomes with distinct function composition. *FEBS Lett.* **389,** 96–101 (1996).
26. Liu, J. & Rost, B. Comparing function and structure between entire proteomes. *Protein Sci.* **10,** 1970–1979 (2001).

27. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA.* **99,** 3695–3700 (2002).
28. Raymond, J. & Segre, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311,** 1764–1767 (2006).
29. Hedges, S. B., Blair, J. E., Venturi, M. L. & Shoe, J. L. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4,** 2 (2004).
30. Martin, W. *et al.* Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. *IUBMB Life* **55,** 193–204 (2003).
31. Douzery, E. J. P., Snell, E. A., Bapteste, E., Delsuc, F. & Philippe, H. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc. Natl Acad. Sci. USA.* **101,** 15386–15391 (2004).
32. Sheridan, P. P., Freeman, K. H. & Brenchley, J. E. Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol. J.* **20,** 1–14 (2003).
33. Hedges, S. B. The origin and evolution of model organisms. *Nature Rev. Genet.* **3,** 838–849 (2002).
34. Anderson, J. S. & Sues, H.-D. (eds) *Major Transitions in Vertebrate Evolution* (Indiana Univ. Press, Bloomington and Indianapolis, in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank J. Anderson for help with the estimates of the time of appearance of the organisms used for this study, and D. Schomburg, R. Wünschiers, D. Bauer, A. Scialpi, M. Koornneef, H. Hillebrand, A. M. Tarchi, P. Bruni, T. Wiehe, B. Haubold and T. Rothery for valuable discussions.

Author Contributions C.A. initiated and devised the project; C.A., S.C. and J.K. analysed the data; and S.C. and C.A. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.A. (Claudia.Acquisti.1@asu.edu).

ARTICLES

Crystal structure of a protein phosphatase 2A heterotrimeric holoenzyme

Uhn Soo Cho[1] & Wenqing Xu[1]

Protein phosphatase 2A (PP2A) is a principal Ser/Thr phosphatase, the deregulation of which is associated with multiple human cancers, Alzheimer's disease and increased susceptibility to pathogen infections. How PP2A is structurally organized and functionally regulated remains unclear. Here we report the crystal structure of an AB′C heterotrimeric PP2A holoenzyme. The structure reveals that the HEAT repeats of the scaffold A subunit form a horseshoe-shaped fold, holding the catalytic C and regulatory B′ subunits together on the same side. The regulatory B′ subunit forms pseudo-HEAT repeats and interacts with the C subunit near the active site, thereby defining substrate specificity. The methylated carboxy-terminal tail of the C subunit interacts with a highly negatively charged region at the interface between A and B′ subunits, suggesting that the C-terminal carboxyl methylation of the C subunit promotes B′ subunit recruitment by neutralizing charge repulsion. Together, our structural results establish a crucial foundation for understanding PP2A assembly, substrate recruitment and regulation.

Dynamic phosphorylation and dephosphorylation of proteins are fundamental mechanisms for cell regulation. Protein phosphatase 2A (PP2A) comprises as much as 1% of total cellular proteins and, together with protein phosphatase 1 (PP1), accounts for >90% of all Ser/Thr phosphatase activities in most tissues and cells[1,2]. A myriad of evidence has demonstrated that PP2A is a critical regulator of many, if not most, aspects of cellular activities[3–6]. Once thought of as a single, broad-specificity phosphatase, PP2A is actually a family of phosphatases that share similar heterotrimeric architecture. The core enzyme of PP2A comprises a ~65-kDa scaffolding A subunit and a ~36-kDa catalytic C subunit. PP2A activities are largely regulated by the binding of one of at least 18 regulatory B subunits to the AC core enzyme, which have been implicated in controlling PP2A substrate specificity, cellular localization and enzymatic activity. The scaffold A subunit is composed of 15 HEAT repeats, whereas the C subunit contains a catalytic domain that shares sequence homology with other Ser/Thr phosphatases such as PP1, PP2B (calcineurin), PP4 and PP6. On the basis of sequence homology, regulatory B subunits can be classified into B (B55), B′ (B56) and B′′ families.

In addition to the catalytic domain that is homologous to PP1, the PP2A C subunit has a unique C-terminal tail (residues 294–309), which contains a motif ($TPDY_{307}FL_{309}$) that is conserved in the catalytic subunits of all PP2A-like phosphatases including PP4 and PP6, and has a critical role in PP2A regulation. There is evidence that methylation of the carboxylate group of the C-terminal residue Leu 309 promotes the recruitment of the regulatory B/B′/B′′ subunit to the AC core dimer[7–11]. Deletion of the C-terminal tail of the C subunit abolishes C-terminal carboxyl methylation and binding of B/B′/B′′ subunits to the AC core complex[7]. The importance of C subunit carboxyl methylation is underlined by the observation of its varied methylation state during the cell cycle and its potential role in Alzheimer's disease pathogenesis[12,13].

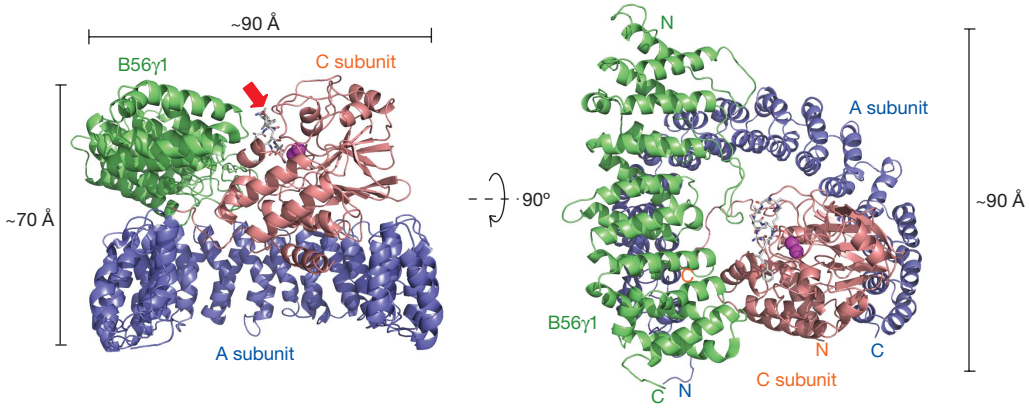


Figure 1 | Overall structure of the Aα–B56γ1–Cα heterotrimeric PP2A holoenzyme. A cartoon illustration of the 'front' and 'top' views of the heterotrimeric PP2A holoenzyme. The scaffold Aα subunit, catalytic Cα subunit and regulatory B56γ1 subunit are coloured in blue, orange and green, respectively. In addition, two metal ions in the active site of the Cα subunit are coloured purple, and microcystin-LR, a PP2A-specific inhibitor, is shown in stick representation. The red arrow points to the position of the active site. The overall size of the trimeric complex is about ~90 Å × ~90 Å × ~70 Å.

[1]Department of Biological Structure, University of Washington, Seattle, Washington 98195, USA.

The B56 (B′) family is the largest B family, consisting of at least eight members. These share a conserved core domain and have critical roles in cell cycle, cell proliferation and Wnt signalling, by dephosphorylating some key regulators of cell activities, including APC, Erk, Akt, paxillin, cyclin G, mdm2 and p53 (refs 14–18). B56γ, as well as Aα and Aβ (the gene mutations of which are associated with multiple cancer forms), are proposed as tumour suppressors[19]. It has been shown that small-molecule PP2A activators are potential anticancer drugs[20]. To provide the structural basis for understanding PP2A function and regulation, we have determined the crystal structure of a heterotrimeric PP2A-B56γ holoenzyme.

Overall architecture

The human PP2A holoenzyme crystal structure reported here contains full-length Aα and Cα subunits, a near full-length B56γ1 subunit and the PP2A inhibitor microcystin (Supplementary Fig. 1). The scaffolding Aα (also called PR65) subunit contains 15 HEAT repeats, forming a horseshoe shape. Each HEAT repeat consists of two antiparallel α-helices connected by an intra-repeat loop, and adjacent repeats are connected by inter-repeat loops. The catalytic Cα subunit forms a compact ellipsoidal structure similar to that of PP1 (ref. 21). Despite a lack of a canonical HEAT repeat sequence motif, the regulatory B subunit fragment B56γ1(30–437) folds like eight pseudo-HEAT repeats (Fig. 1). Both C and B56 subunits dock on the 'apical' side of the horseshoe-shaped scaffold through interactions with the intra-repeat loops of the A subunit HEAT repeats. The active site of the C subunit faces away from the horseshoe-shaped scaffold and appears to be open for substrate access (Fig. 1). The C and B56 subunits also form extensive interactions. In addition to these interactions between globular structures, the critical C-terminal tail of the C subunit docks on the interface of the A and B subunits, where it could regulate the recruitment of the B subunit (Fig. 1).

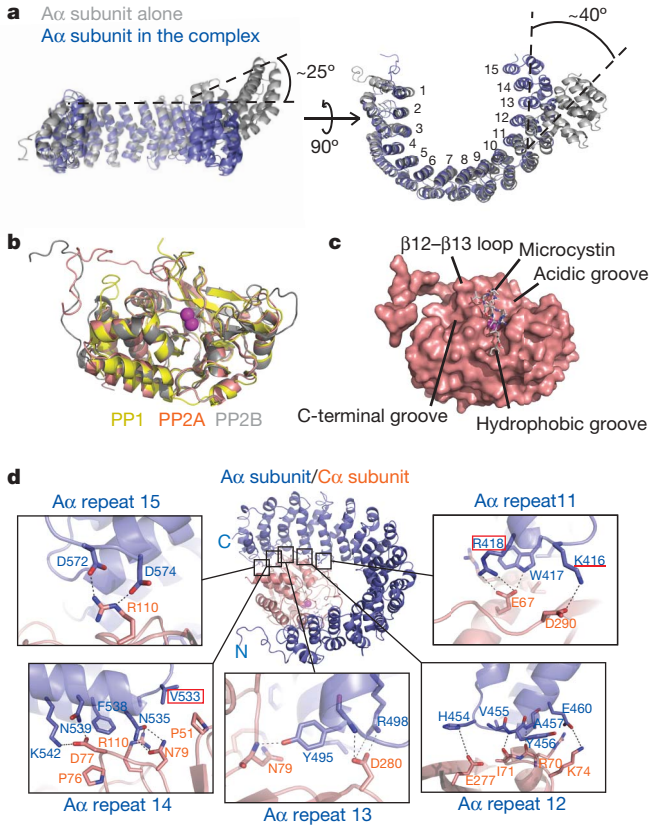


Figure 2 | Structures of the scaffold A subunit, the catalytic Cα subunit and the A–C interface. **a**, Structure comparison of the Aα subunit alone (grey) and in the trimeric complex (blue), with front and top views. In the complex structure, the A subunit intra-repeat loops, which are responsible for B and C subunit binding, form a circle narrower than that formed by inter-repeat loops. Conformational changes of repeats 11–15 are mostly derived from the interface change between repeats 12–13. **b**, A structure superposition of the PP2A Cα subunit (orange), PP1 (yellow) and PP2B catalytic domain (grey). **c**, A surface representation of the Cα subunit with microcystin-LR (shown in stick representation). **d**, Interface of the Aα (blue) and Cα subunits (orange). Red boxes represent positions of genetic mutations found in melanoma cancer (R418W)[27] and colon adenocarcinoma (V545A in PP2A Aβ)[29]; the red underline indicates the position of a previously identified mutant that interrupts the A–C interaction (K416E)[28]. Dashed lines indicate hydrogen bonds.

A and C subunits and the A–C interface

Like the previously reported monomeric A subunit structure[22], the A subunit in the PP2A trimeric complex also contains 15 HEAT repeats. However, there is a substantial conformational change from the twisted hook shape of the monomeric structure to a more closed horseshoe shape of the trimeric structure of the A subunit (Fig. 2a). This overall conformational change is mostly observed between HEAT repeat 11 and repeat 15, which are involved in C-subunit binding. Specifically, this conformational change is largely derived from HEAT repeat 12, the only repeat that does not show a kink at the helix centre in the monomeric A subunit structure. In the PP2A AB′C complex, presumably owing to either the binding of C subunit or the binding of B56γ1 with the AC dimeric complex, the packing of repeat 12 is rearranged into a similar conformation with other HEAT repeats in the complex (Supplementary Fig. 2): the previously seen left-handed 45° rotation of repeat 13 relative to 12 was changed to 20°. Future studies will be needed to understand whether this conformational change has a regulatory role in PP2A assembly.

The PP2A Cα subunit shares sequence homology and not surprisingly shows high similarity in three-dimensional structure to the catalytic subunits of PP1 and PP2B (calcineurin), with Cα root-mean-square (r.m.s.) differences of 1.5 Å (from 284 Cα subunits) and 1.8 Å (from 275 Cα subunits), respectively (Fig. 2b)[21,23]. Two

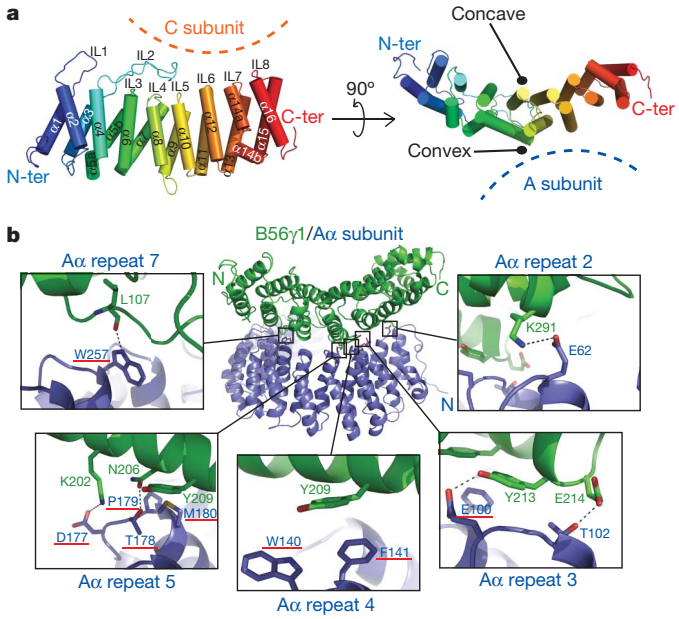


Figure 3 | Overall structure of the regulatory B56γ1 subunit and the A–B interface. **a**, Front and bottom views of the overall structure of the regulatory B56γ1 subunit. The peptide chain is colour-coded from blue to red, from the N terminus to the C terminus. Helices are labelled from α1 to α16, and each of helices 5 and 14 contains two short helices (a and b). Intra-repeat loops of each pseudo-HEAT repeat are labelled IL1–IL8. The binding surfaces for the A and C subunits are also shown schematically. **b**, Interface of the Aα (blue) and B56γ1 subunits (green). Positions of mutations known to disrupt A–B interaction are underlined in red[30].

metal ions (presumably Mn^{2+})[24] and the PP2A-specific inhibitor microcystin bind to the C subunit in a manner almost identical to that of PP1 (ref. 21) (Fig. 2c). Most residues involved in catalysis are highly conserved, and it is likely that PP2A deploys a catalytic mechanism similar to that of PP1 and PP2B. Despite almost identical active sites, there is substantial surface variation among the catalytic subunits of PP2A, PP1 and other Ser/Thr phosphatases that determines distinct subunit assembly properties and contributes to substrate specificity (Supplementary Figs 3 and 4).

The A and C subunits of PP2A can form a stable AC core dimer[25,26]. In our structure, the A subunit makes extensive interactions with the C subunit through intra-repeat loops and inner helices of the A subunit repeats 11–15. We observed a network of hydrogen bonding, ionic and hydrophobic interactions in the A–C interface. Detailed interactions are shown in Fig. 2d. An R418W mutation was found in melanoma cancer patients that disrupts the A–C interaction[27]. In addition, a K416E mutation was also reported to abolish the A–C interaction[28]. In our crystal structure, both R418 and K416 are located in the A–C interface and form specific hydrogen bonds with the C subunit (Fig. 2d). Another tumorigenic mutation found in Aβ, which corresponds to V533A in Aα, is also located in the A–C interface[29]. Therefore, our structure provides a molecular basis for cancer-associated, A-subunit mutations that disrupt PP2A assembly.

B56 subunit and the A–B56 interface

All B56/B′ family members contain a central conserved domain with 80% sequence identity (Supplementary Fig. 5). In the crystal structure, B56γ1(30–437) covers the entire conserved domain and is sufficient for the formation of the stable AB′C heterotrimeric complex (Supplementary Figs 1 and 5). The structure of B56γ1(30–437) contains eight two-helix units that stack to form a solenoid-shaped structure in a manner similar to that of HEAT repeats, despite a lack of sequence homology between B56 and any HEAT repeat protein. For convenience, we describe the B56γ1(30–437) structure as comprising eight pseudo-HEAT repeats, with helices 5 and 14 (in repeats 3 and 7, respectively) containing two shorter helices (Fig. 3a). Repeats 1 and 2 contain unusually long intra-repeat loops, and intra-repeat loop 2 is involved in C subunit binding (see below). There is no apparent internal repetitive sequence motif among these eight pseudo-HEAT repeats. The last 31 residues of B56γ1(30–437) are not visible in our electron density map, probably because of structural flexibility in this region.

The A–B56 interface is formed mostly between the intra-repeat loops of A-subunit HEAT repeats 2–7 and the convex side of the B56 subunit pseudo-HEAT repeats. Whereas the A subunit uses intra-repeat loops on the apical side for binding, the B56 subunit uses its C-terminal half of the convex helices of pseudo-HEAT repeats for interactions (Fig. 3a). Detailed interactions are shown in Fig. 3b. Overall, the interface between A and B56 subunits is relatively loose, which is consistent with the fact that A and B subunits cannot directly form a stable complex. The weak binding between A and B subunits is enhanced by binding of the methylated C-terminal tail to this interface (see below). The significance of the interface residues revealed by the crystal structure is supported by previous mutagenesis studies[30]. Notably, all of the A subunit residues shown to be important for A–B56 interactions are found in the A–B56 interface in our crystal structure (Fig. 3b).

B56–C interface and substrate specificity

The interface between B56 and C subunits consists of three discrete areas (Fig. 4a). The first interface is formed between a conserved arginine residue (Arg 268) in the C subunit β12–β13 loop and an acidic cluster in the long intra-repeat loop 2 of B56γ1. The second interface is formed between the helical subdomain of the C subunit and intra-repeat loops of B56γ1 HEAT repeats 6–8. The third interface is formed between the C-terminal tail and B56γ1 HEAT repeats 4–6.

Previous studies have demonstrated that, to a large extent, the regulatory B subunits control PP2A substrate specificities. From a structural point of view, the B subunit can provide PP2A specificity by shielding a substantial surface of the PP2A C subunit and, at the same time, provide a new surface near the catalytic active site. In this regard, the large concave surface of the B56γ1 HEAT repeats is brought into close proximity to the C subunit active site (Fig. 4b) and may provide the docking site for some of the AB′C PP2A substrates. In addition to this concave surface, more N- and C-terminal regions of B56 may also be involved in substrate recruitment, which is exemplified by mutation studies of B56(Δγ1) (Fig. 4a; see also Supplementary Figs 6 and 7)[16]. Overall, the formation of the BC complex results in a completely different physicochemical landscape

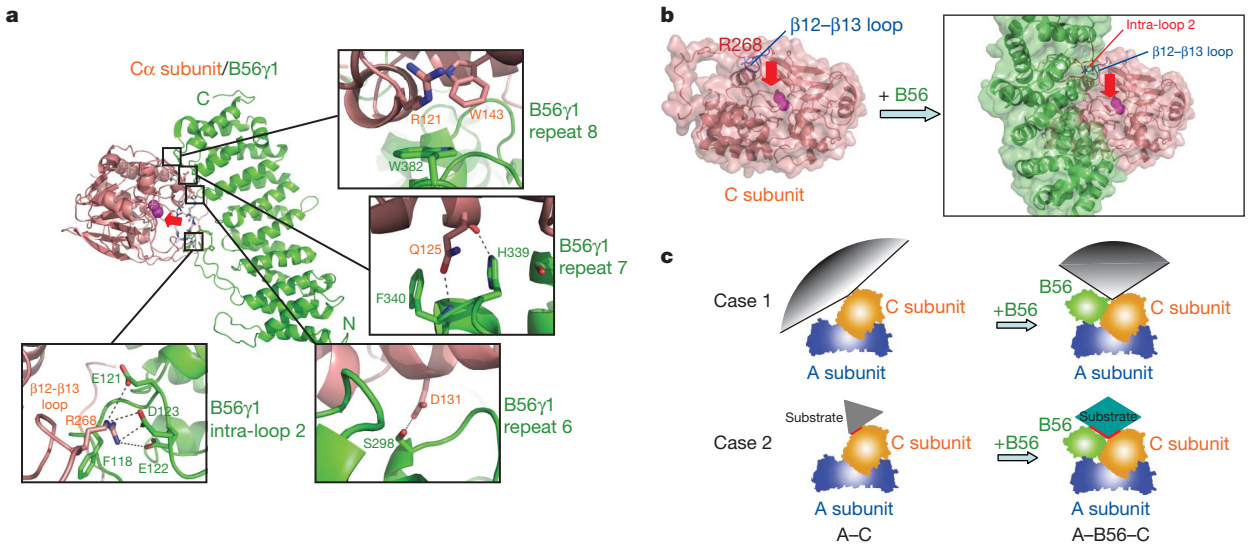


Figure 4 | Interface of the catalytic Cα subunit with the regulatory B56γ1. **a**, Interface of the Cα subunit (orange) and B56γ1 (green). The red arrow points to the position of the active site in the Cα subunit. **b**, A surface representation of the Cα subunit with and without B56γ1 bound. The β12–β13 loop of the Cα subunit interacts with intra-repeat loop 2 of B56γ1 through salt bridges between R268 in the Cα subunit and E121, E122 and D123 in B56γ1. **c**, A schematic model of B56-binding-induced substrate specificity change. The binding of the B56 subunit markedly changes the environment near the active site, by both limiting the accessible surface to the active site (case 1) and providing novel potential substrate binding surfaces (case 2).

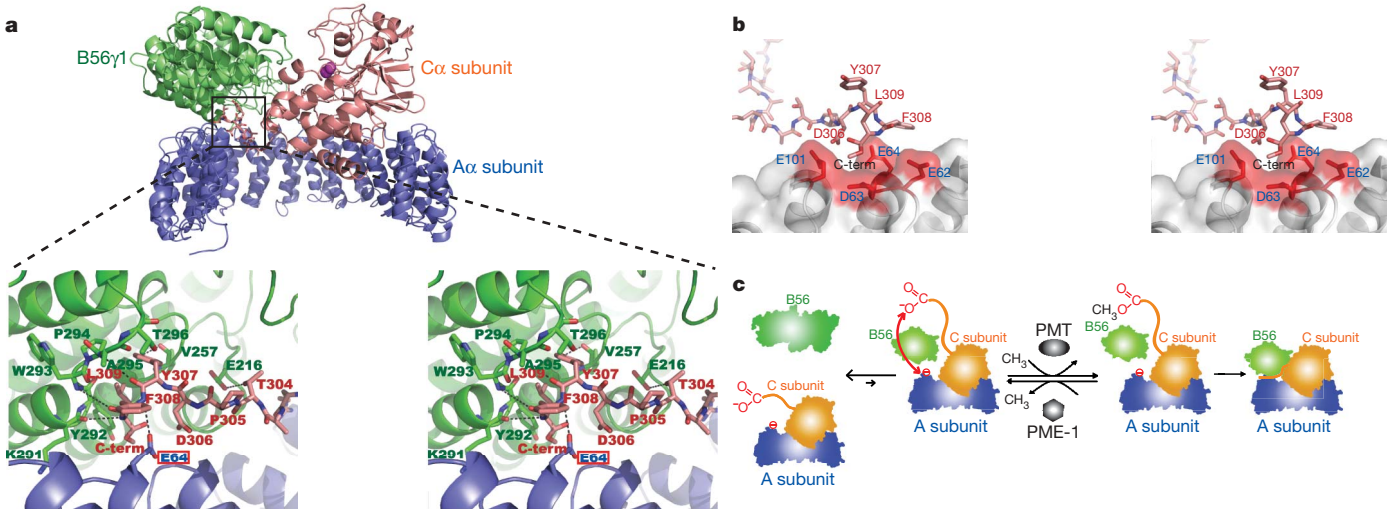


Figure 5 | Interactions between the methylated C-terminal tail of the catalytic Cα subunit and the A–B interface. **a**, Overall position of the C-terminal tail of the Cα subunit in the trimeric complex and a stereo view of detailed interactions. The red box indicates the position of genetic mutations E64D and E64G found in lung and breast cancer patients, respectively. **b**, The methylated C-terminal tail of the Cα subunit is positioned in the middle of four negatively charged residues of the Aα subunit (in red). **c**, A schematic model of the regulatory mechanism of C-terminal tail methylation in formation of the trimeric A–B56–C complex. The affinity between the B56 subunit and the AC core enzyme is too low to form a stable complex, and the unmethylated C-terminal tail of the C subunit cannot settle down on the A–B interface due to the charge repulsion between the carboxyl group of C-terminal Leu 309 and the negatively charged surface on the A subunit near the proposed binding area. When the C-terminal tail of the C subunit is methylated by phosphatase methyltransferase (PMT), charge repulsion is neutralized and the methylated C-terminal tail can settle down on the A–B interface. This allows for the recruitment of B56 to form a stable AB′C heterotrimeric complex.

of the active site compared with the C subunit alone, which determines the substrate specificity, and different substrates can be potentially recruited via different surface areas of the BC complex (Fig. 4c).

PP2A regulation by C-terminal methylation

It is well established that methylation of the C-terminal carboxyl group of the C subunit controls recruitment of the regulatory B subunit to the AC core enzyme. Consistent with previously reported results[10,31], the C-terminal carboxyl group of our PP2A Cα subunit is methylated in the baculovirus insect cell expression system, as confirmed by western blot using a carboxyl-methylated PP2A-specific antibody (data not shown). The C-terminal tail mostly interacts with HEAT repeats 4–6 of the B subunit. The only hydrogen bond between the C-terminal tail and the A subunit is formed between the main-chain amine group of the C subunit residue Phe 308 and the side chain of the A subunit residue Glu 64 (Fig. 5a, b). It should be noted that the E64D and E64G mutations have been found in lung and breast cancer patients, respectively[27]. Either mutation would disrupt the observed hydrogen bond between Glu 64 and the C-terminal tail.

Notably, the C-terminal carboxyl-methylated residue Leu 309 is located in the centre of a highly negatively charged environment formed by residues Glu 62, Asp 63, Glu 64 and Glu 101 of the A subunit and Asp 306 of the C subunit (Fig. 5a, b). In the absence of C-terminal carboxyl methylation, the charge–charge repulsion between the negatively charged C-terminal carboxyl group and the acidic cluster would not favour the docking of the C-terminal tail in this area. We propose that the neutralization of the charge–charge repulsion by methylation would allow docking of the C-terminal tail and promote the binding of the B56 subunit to the AC core enzyme (Fig. 5c). Future studies will be needed to examine whether the B55 family is also regulated with a mechanism similar to that of the B56 (B′) family. Nevertheless, a similar mechanism is used by Ras, in which carboxyl methylation of the terminal cysteine residue favours membrane attachment by converting the hydrophilic tail to a hydrophobic one, although methylation only has an auxiliary role in Ras membrane attachment[32].

PP2A is also regulated by phosphorylation. For example, phosphorylation of the C subunit residue Tyr 307 by tyrosine kinases such as Src inhibits PP2A activity[33], and phosphorylation of the B56 subunit by Erk inhibits PP2A assembly[15]. Our work provides the structural basis for PP2A regulation by phosphorylation (Supplementary Figs 8 and 9)[15].

Our crystal structure reveals the overall architecture of the AB′C family PP2A holoenzyme, which provides the structural basis for understanding PP2A holoenzyme assembly, substrate recruitment and regulation by phosphorylation and methylation.

METHODS

Protein purification and crystallization. PP2A holoenzyme heterotrimers were reconstituted by mixing highly purified A, B and C subunits, further purified, and used for extensive screening of crystallization conditions. Purification and reconstitution are described in more detail in Supplementary Information. The human Aα–B56γ1–Cα heterotrimeric complex was concentrated to ~8 mg ml^{-1} with a threefold molar excess of microcystin-LR (MCLR), a specific PP2A inhibitor. The PP2A Aα–B56γ1–Cα–microcystin complex was crystallized in 0.1 M MES (pH 7.0), 50 mM NaCl, 1.4 M ammonium sulphate, 0.2 M LiCl, 2% 1,6-diaminohexane. Crystals were cryo-protected with 18% glycerol and frozen with liquid nitrogen. The space group of this crystal form is $P2_13$, and unit cell dimensions are $a = b = c = 265$ Å, $\alpha = \beta = \gamma = 90°$.

Crystal structure determination. The structure was determined by the single anomalous dispersion method (SAD) using Se-Met-substituted A subunit or both A and B subunits. Two SAD data sets of A$^{Se\text{-}Met}$–B56–C complex crystals were collected with 3.7 Å and 3.5 Å resolutions at peak wavelength, with and without inverted beam diffraction, respectively. These data sets were integrated and scaled using HKL2000 (ref. 34). There were two heterotrimeric complexes in each asymmetric unit with 76% solvent content. We found 6 out of 28 possible Se sites using the 3.7 Å SAD data set with the shake-and-bake algorithm[35]. These sites were refined and used to find 19 more sites using SHARP[36]. Phase extension and solvent flattening were performed with a 3.5 Å SAD data set by DM in the CCP4 package[37]. Owing to the high solvent content, the number of diffractions in our 3.5 Å resolution data set is equivalent to that of a ~3.1 Å resolution data set with a 50% solvent content. As a result, the initial electron density map from SAD phasing and solvent flattening was of excellent quality and allowed us to trace the entire peptide chain for all three subunits without any ambiguity (Supplementary Fig. 1). The initial structural model was built with Xtalview[38] and COOT[39]. Refmac5 (ref. 40) was used for the TLS and restrained refinement. After model building of B56γ1, positions of Met residues in the structural model were confirmed by Se anomalous difference map with a 3.75 Å A$^{Se\text{-}Met}$–B56$^{Se\text{-}Met}$–C SAD data set. The final model has a working *R*-factor of 25.6%

and a free *R*-factor of 31.6% (Supplementary Table 1). None of the non-glycine residues is in the disallowed region of the Ramachandran plot. All figures were generated by Pymol (http://pymol.sourceforge.net/).

Received 11 September; accepted 16 October 2006.
Published online 1 November 2006.

1. Lin, X. H. *et al.* Protein phosphatase 2A is required for the initiation of chromosomal DNA replication. *Proc. Natl Acad. Sci. USA* **95,** 14693–14698 (1998).
2. Depaoli-Roach, A. A. *et al.* Serine/threonine protein phosphatases in the control of cell function. *Adv. Enzyme Regul.* **34,** 199–224 (1994).
3. Sontag, E. Protein phosphatase 2A: the Trojan Horse of cellular signaling. *Cell. Signal.* **13,** 7–16 (2001).
4. Janssens, V. & Goris, J. Protein phosphatase 2A: a highly regulated family of serine/threonine phosphatases implicated in cell growth and signalling. *Biochem. J.* **353,** 417–439 (2001).
5. Goldberg, Y. Protein phosphatase 2A: who shall regulate the regulator? *Biochem. Pharmacol.* **57,** 321–328 (1999).
6. Virshup, D. M. Protein phosphatase 2A: a panoply of enzymes. *Curr. Opin. Cell Biol.* **12,** 180–185 (2000).
7. Ogris, E., Gibson, D. M. & Pallas, D. C. Protein phosphatase 2A subunit assembly: the catalytic subunit carboxy terminus is important for binding cellular B subunit but not polyomavirus middle tumor antigen. *Oncogene* **15,** 911–917 (1997).
8. Tolstykh, T., Lee, J., Vafai, S. & Stock, J. B. Carboxyl methylation regulates phosphoprotein phosphatase 2A by controlling the association of regulatory B subunits. *EMBO J.* **19,** 5682–5691 (2000).
9. Wu, J. *et al.* Carboxyl methylation of the phosphoprotein phosphatase 2A catalytic subunit promotes its functional association with regulatory subunits *in vivo*. *EMBO J.* **19,** 5672–5681 (2000).
10. Yu, X. X. *et al.* Methylation of the protein phosphatase 2A catalytic subunit is essential for association of Bα regulatory subunit but not SG2NA, striatin, or polyomavirus middle tumor antigen. *Mol. Biol. Cell* **12,** 185–199 (2001).
11. Wei, H. *et al.* Carboxymethylation of the PP2A catalytic subunit in *Saccharomyces cerevisiae* is required for efficient interaction with the B-type subunits Cdc55p and Rts1p. *J. Biol. Chem.* **276,** 1570–1577 (2001).
12. Turowski, P., Fernandez, A., Favre, B., Lamb, N. J. & Hemmings, B. A. Differential methylation and altered conformation of cytoplasmic and nuclear forms of protein phosphatase 2A during cell cycle progression. *J. Cell Biol.* **129,** 397–410 (1995).
13. Sontag, E. *et al.* Downregulation of protein phosphatase 2A carboxyl methylation and methyltransferase may contribute to Alzheimer disease pathogenesis. *J. Neuropathol. Exp. Neurol.* **63,** 1080–1091 (2004).
14. Janssens, V., Goris, J. & Van Hoof, C. PP2A: the expected tumor suppressor. *Curr. Opin. Genet. Dev.* **15,** 34–41 (2005).
15. Letourneux, C., Rocher, G. & Porteu, F. B56-containing PP2A dephosphorylate ERK and their activity is controlled by the early gene IEX-1 and ERK. *EMBO J.* **25,** 727–738 (2006).
16. Ito, A. *et al.* A truncated isoform of the PP2A B56 subunit promotes cell motility through paxillin phosphorylation. *EMBO J.* **19,** 562–571 (2000).
17. Seeling, J. M. *et al.* Regulation of β-catenin signaling by the B56 subunit of protein phosphatase 2A. *Science* **283,** 2089–2091 (1999).
18. McCright, B. & Virshup, D. M. Identification of a new family of protein phosphatase 2A regulatory subunits. *J. Biol. Chem.* **270,** 26123–26128 (1995).
19. Chen, W. *et al.* Identification of specific PP2A complexes involved in human cell transformation. *Cancer Cell* **5,** 127–136 (2004).
20. Perrotti, D. & Neviani, P. ReSETting PP2A tumour suppressor activity in blast crisis and imatinib-resistant chronic myelogenous leukaemia. *Br. J. Cancer* **95,** 775–781 (2006).
21. Goldberg, J. *et al.* Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* **376,** 745–753 (1995).
22. Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A. & Barford, D. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* **96,** 99–110 (1999).
23. Griffith, J. P. *et al.* X-ray structure of calcineurin inhibited by the immunophilin-immunosuppressant FKBP12–FK506 complex. *Cell* **82,** 507–522 (1995).
24. Cai, L., Chu, Y., Wilson, S. E. & Schlender, K. K. A metal-dependent form of protein phosphatase 2A. *Biochem. Biophys. Res. Commun.* **208,** 274–279 (1995).
25. Kremmer, E., Ohst, K., Kiefer, J., Brewis, N. & Walter, G. Separation of PP2A core enzyme and holoenzyme with monoclonal antibodies against the regulatory A subunit: abundant expression of both forms in cells. *Mol. Cell. Biol.* **17,** 1692–1701 (1997).
26. Ruediger, R. *et al.* Identification of binding sites on the regulatory A subunit of protein phosphatase 2A for the catalytic C subunit and for tumor antigens of simian virus 40 and polyomavirus. *Mol. Cell. Biol.* **12,** 4872–4882 (1992).
27. Ruediger, R., Pham, H. T. & Walter, G. Disruption of protein phosphatase 2A subunit interaction in human cancers with mutations in the Aα subunit gene. *Oncogene* **20,** 10–15 (2001).
28. Turowski, P., Favre, B., Campbell, K. S., Lamb, N. J. & Hemmings, B. A. Modulation of the enzymatic properties of protein phosphatase 2A catalytic subunit by the recombinant 65-kDa regulatory subunit PR65α. *Eur. J. Biochem.* **248,** 200–208 (1997).
29. Wang, S. S. *et al.* Alterations of the PPP2R1B gene in human lung and colon cancer. *Science* **282,** 284–287 (1998).
30. Ruediger, R., Fields, K. & Walter, G. Binding specificity of protein phosphatase 2A core enzyme for regulatory B subunits and T antigens. *J. Virol.* **73,** 839–842 (1999).
31. Ikehara, T., Shinjo, F., Ikehara, S., Imamura, S. & Yasumoto, T. Baculovirus expression, purification, and characterization of human protein phosphatase 2A catalytic subunits α and β. *Protein Expr. Purif.* **45,** 150–156 (2006).
32. Philips, M. R. Methotrexate and Ras methylation: a new trick for an old drug? *Sci. STKE* **2004,** pe13 (2004).
33. Chen, J., Martin, B. L. & Brautigan, D. L. Regulation of protein serine-threonine phosphatase type-2A by tyrosine phosphorylation. *Science* **257,** 1261–1264 (1992).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
35. Smith, G. D. *et al.* The use of SnB to determine an anomalous scattering substructure. *Acta Crystallogr. D* **54,** 799–804 (1998).
36. De La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
37. Cowtan, K. DM: an automated procedure for phase improvement by density modification. *Joint CCP4 ESF-EACBM Newsl. Prot. Crystallogr.* **31,** 34–38 (1994).
38. McRee, D. E. XtalView/Xfit–A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125,** 156–165 (1999).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
40. Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr. D* **55,** 247–255 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Morrone, F. Gao and other laboratory members and rotation students for help with this work. We are grateful to J. Abendroth for advice on crystallographic computation, and the staff at ALS beamline 5.0.2 for assistance with data collection. We also thank D. Virshup and X. Liu for B56 cDNAs, and N. Zheng, D. Virshup and E. Ogris for critical comments on this manuscript. This work was supported in part by an Investigator's Award from the Burroughs Welcome Fund to W.X. and by the Keck Center for Pathogenesis at the University of Washington.

Author Information Coordinates and structure factors are deposited in the Protein Data Bank under accession code 2IAE. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.X. (wxu@u.washington.edu).

LETTERS

Pulsar spins from an instability in the accretion shock of supernovae

John M. Blondin[1] & Anthony Mezzacappa[2]

Rotation-powered radio pulsars are born with inferred initial rotation periods[1] of order 300 ms (some as short as 20 ms) in core-collapse supernovae. In the traditional picture, this fast rotation is the result of conservation of angular momentum during the collapse of a rotating stellar core. This leads to the inevitable conclusion that pulsar spin is directly correlated with the rotation of the progenitor star[2]. So far, however, stellar theory has not been able to explain the distribution of pulsar spins, suggesting that the birth rotation is either too slow[3] or too fast[2,4]. Here we report a robust instability of the stalled accretion shock in core-collapse supernovae that is able to generate a strong rotational flow in the vicinity of the accreting proto-neutron star. Sufficient angular momentum is deposited on the proto-neutron star to generate a final spin period consistent with observations, even beginning with spherically symmetrical initial conditions. This provides a new mechanism for the generation of neutron star spin and weakens, if not breaks, the assumed correlation between the rotational periods of supernova progenitor cores and pulsar spin.

The collapse of a massive star's core that triggers a supernova explosion is followed by a brief epoch of less than a second during which the nascent supernova shock wave stalls at a radius of order 100 km and is revived, and the supernova initiated, by an as yet undetermined mechanism[5]. Hydrodynamics simulations have shown that this quasi-steady shock is subject to the stationary accretion shock instability, or SASI[6–9]. However, these two-dimensional simulations admit only axisymmetric modes and, hence, the resulting dynamics cannot affect the rotation of the accretion flow. As we show here, in three dimensions non-axisymmetric modes can significantly alter the angular momentum of the collapsed core.

We have performed on a three-dimensional cartesian grid a series of simulations of a steady accretion shock, following the numerical approach described in ref. 6 and in the Supplementary Information. We found that the nonlinear evolution of the SASI is dominated by a low-order non-axisymmetric mode characterized by a spiral flow pattern beneath the accretion shock. The SASI has been interpreted in terms of a growing acoustic wave propagating around the periphery of the shocked accretion flow—that is, around the periphery of the region between the proto-neutron star (PNS) and accretion shock, or post-shock region[10]. Whereas the axisymmetric sloshing mode (characterized by $l = 1$ in spherical harmonics) seen in earlier work represents the propagation of this wave along a symmetry axis from one pole to the other, the spiral mode ($m = 1$) discovered here represents the propagation of this wave around an axis, as illustrated in Fig. 1.

In the nonlinear regime this SASI wave propagating around the inside surface of the accretion shock creates two strong counter-rotating flows as seen in Fig. 2. Over the leading half of the SASI wave, the gas immediately behind the accretion shock is moving in the same direction as the propagation of the wave. This flow is fed in part by the obliquity of the accretion shock, which refracts the radially falling gas above the shock into a rotational motion moving with the SASI wave. As the leading edge of the SASI wave travels around the accretion shock, it subverts the weaker, receding portion of the accretion shock ahead of it and drives the lower-entropy shocked gas interior to the receding portion of the shock down towards the PNS. The orientation of the weak accretion shock in this region leads to a post-shock flow moving in a direction opposite to that of the SASI wave.

The spiral mode of the SASI is a robust result of a stalled accretion shock in three dimensions. We have evolved a dozen simulations with different initial perturbations, from random acoustic noise in the accretion shock cavity to various configurations of density perturbations in the infalling material above the shock. In all cases the late-time evolution was dominated by this spiral mode. We also ran three simulations (using different initial perturbations) with moderate rotation of the infalling gas to model the effect of a rotating progenitor star, using a specific angular momentum at the accretion shock comparable to the 15-solar-mass model described in ref. 4. In

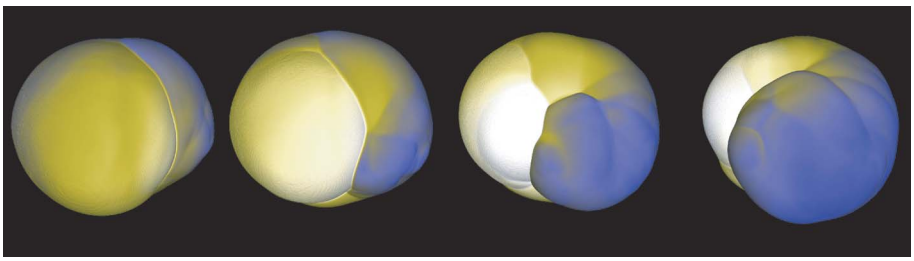


Figure 1 | The evolution of the supernova accretion shock illustrates the rotation of the spiral mode of the SASI. The blue portion of the shock surface represents the leading portion of the spiral SASI wave, seen here propagating from right to left across the front face of the shock. The discontinuity between the blue and white surfaces is the shock triple point marking the leading edge of the SASI wave. An animation of this evolution is available in Supplementary Information.

[1]Department of Physics, North Carolina State University, Raleigh, North Carolina 27695-8202, USA. [2]Physics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6354, USA.

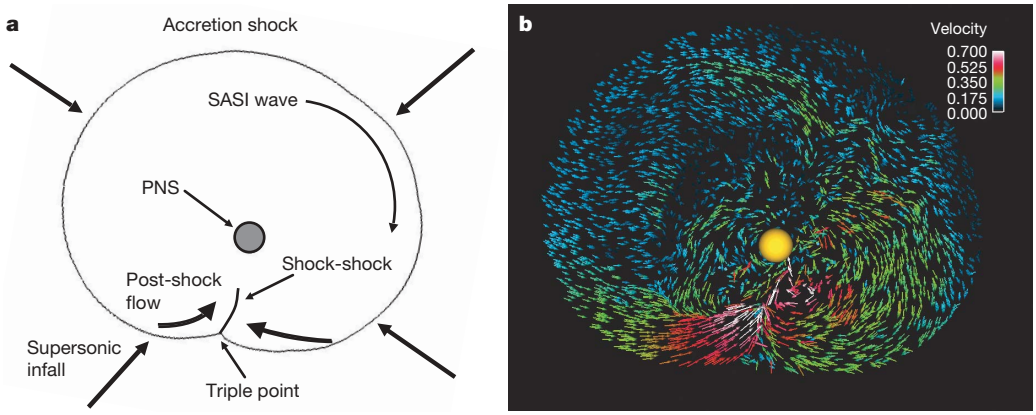


Figure 2 | The flow in the equatorial plane of the spiral SASI mode drives accretion of angular momentum onto the PNS. a, This diagram illustrates the shock structure and corresponding post-shock accretion flow created by the spiral SASI wave. The location of the accretion shock is taken from the equatorial plane of a three-dimensional simulation with the shock pattern (the SASI wave) propagating in a clockwise direction. The leading edge of the internal SASI wave is marked by a shock-shock[13]: a shock wave formed by the steepening of a pressure wave propagating along the inside surface of the accretion shock. This shock-shock connects to the accretion shock at a triple point, seen as a discontinuity in the surface of the accretion shock. In three dimensions this triple point is a line segment on the surface of the accretion shock that spans roughly half the circumference, as seen in Fig. 1. **b**, The flow vectors highlight two strong rotational flows. On the right the flow is moving clockwise along with the shock pattern, whereas at the bottom left the post-shock flow is being diverted into a narrow stream moving anticlockwise, fuelling the accretion of angular momentum onto the PNS.

the presence of progenitor rotation, the spiral SASI becomes dominant much more quickly than in the absence of rotation.

The spiral flow pattern generated by the distorted accretion shock will have a marked effect on the underlying PNS. Figure 3 shows the time evolution of the net angular momentum accreted onto the PNS. For the non-rotating-progenitor models there is no angular momentum in the flow entering the simulation domain at the outer boundary, nor are there any external torques that might change the global angular momentum. Therefore the net angular momentum of the simulation must remain zero. As a consequence, the angular momentum in the accretion flow above the surface of the PNS should be equal and opposite to the angular momentum of the accreting PNS. Our simulations maintain this equality to within a few per cent, with the difference attributable to numerical errors inherent in advecting angular momentum on a cartesian grid. The separation of angular momentum plotted in Fig. 3 is a direct result of the spiral SASI wave, which generates two counter-rotating flows as described above. The net result is that gas with angular momentum of one sign is accreted onto the PNS, whereas gas with the opposite angular momentum flows around the outer regions of the shock cavity, presumably driven outwards once the supernova explosion is initiated. The most striking feature in Fig. 3 is the almost linear increase in the PNS angular momentum about 1 s after the simulation begins, and even earlier for a rotating progenitor.

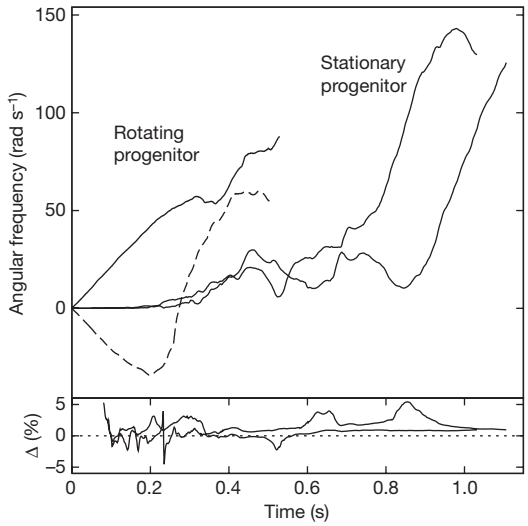


Figure 3 | The spin-up of the accreting PNS due to the spiral SASI mode. The total accreted angular momentum (divided by the moment of inertia of an isolated neutron star such that the unit of measure is the spin frequency of the relic neutron star) is plotted as a function of time for three three-dimensional SASI simulations with different initial perturbations. The bottom panel shows the difference between the angular momentum of the PNS and the total angular momentum in the flow above the surface of the PNS for the non-rotating-progenitor models. Their vector sum should, in fact, cancel given that we began with spherically symmetrical initial conditions and, hence, no net angular momentum. For the rotating-progenitor model we also show the x component of the angular momentum (dashed line). At early times the accreted angular momentum is dominated by the progenitor rotation (aligned along the $-x$ axis), but once the SASI becomes nonlinear it dominates the angular momentum accretion rate, and the rotation of the accreted material reverses direction.

The presence of rotation in the infalling gas helps to initiate the spiral mode of the SASI, as demonstrated by the early increase in angular momentum seen in Fig. 3 and the fact that the rotation axis of the SASI wave is roughly aligned (by 10°, 15° and 45°) with the spin axis of the progenitor star. This alignment has the surprising effect of erasing the progenitor spin from the PNS. Because the angular momentum accretion driven by the SASI is opposite to that of the SASI wave itself, the initial effect is to spin down the PNS. The comparable rate of increase in accreted angular momentum between the non-rotating and rotating models shows that the magnitude of the angular momentum accretion rate is set by the flow pattern of the spiral SASI wave, not by the angular momentum of the infalling gas above the accretion shock.

The net angular momentum accreted onto the PNS as a result of the spiral SASI mode during the supernova initiation phase can markedly alter the spin rate of the neutron star left behind. Over a timespan of 250 ms the PNS in our models will typically accrete 0.1 solar masses and a net angular momentum of 2.5×10^{47} g cm^2 s^{-1}. If we divide this total angular momentum by the moment of inertia of an isolated neutron star[11], $I \approx 2 \times 10^{45}$, we find that the resulting rotational period would be about 50 ms. (We use the results from ref. 12 to scale our models to relevant supernova values: a shock radius of 230 km, 1.2 solar masses interior to the shock, and a mass accretion rate of 0.36 solar masses per second.) This does not include any relic angular momentum from the progenitor core rotation. Thus, for an initially non-rotating or slowly (angular momentum about 2.5×10^{47} g cm^2 s^{-1} or less) rotating progenitor the SASI will be the dominant source of angular momentum in the remnant neutron star

and will produce a spin rate consistent with the inferred birth periods of pulsars[1]. By comparison, current stellar evolution models[4] predict a relic angular momentum of at least 8×10^{47} g cm^2 s^{-1}, which would result in a birth period of about 15 ms. Such fast rotation is marginally consistent with only the fastest radio pulsars. Either current predictions of progenitor core rotation are too large or some unknown mechanism would serve to spin down the PNS (the SASI would reduce the relic angular momentum by only about 30%). If the former is true and core rotation rates are actually significantly lower, the final spin period of the neutron star would be determined by the SASI, as discussed above.

The precise final spin period of the remnant neutron star will be determined by the length of time for which the SASI spiral mode is dominant, which will depend both on how fast it is initiated after core bounce and how long the stalled accretion shock persists before an explosion is initiated. The former is affected by the (poorly known) progenitor rotation, whereas the latter will require three-dimensional supernova models with sufficient realism to follow the entire explosion process—for example with three-dimensional, multifrequency neutrino transport (the models considered here are valid only for the stalled-shock phase). The different simulations shown in Fig. 3 illustrate the uncertainty in our models of the time at which the spiral flow pattern begins but at the same time confirm the robustness of the spin-up induced by the spiral SASI. The outcomes provide a new mechanism for the generation of neutron star spin. Moreover, they demonstrate that progenitor spin and neutron star spin will not be as simply correlated as previously believed.

Received 7 April; accepted 6 November 2006.

1. Faucher-Giguere, C.-A. & Kaspi, V. M. Birth and evolution of isolated radio pulsars. *Astrophys. J.* **643,** 332–355 (2006).
2. Ott, C. D., Burrows, A., Thompson, T. A., Livne, E. & Walder, R. The spin periods and rotational profiles of neutron stars at birth. *Astrophys. J.* **164** (suppl.), 130–155 (2006).
3. Spruit, H. C. & Phinney, E. S. Birth kicks as the origin of pulsar rotation. *Nature* **393,** 139–141 (1998).
4. Heger, A., Woosley, S. E. & Spruit, H. C. Presupernova evolution of differentially rotating massive stars including magnetic fields. *Astrophys. J.* **626,** 350–363 (2005).
5. Mezzacappa, A. Ascertaining the core collapse supernova mechanism. *Annu. Rev. Nucl. Part. Sci.* **55,** 467–515 (2005).
6. Blondin, J. M., Mezzacappa, A. & DeMarino, C. Stability of standing accretion shocks, with an eye toward core-collapse supernovae. *Astrophys. J.* **584,** 971–980 (2003).
7. Janka, H.-T. h. *et al.* Neutrino-driven supernovae: an accretion instability in a nuclear physics controlled environment. *Nucl. Phys. A* **758,** 19–26 (2004).
8. Scheck, L., Plewa, T., Janka, H.-T. h., Kifonidis, K. & Mueller, E. Pulsar recoil by large-scale anisotropies in supernova explosions. *Phys. Rev. Lett.* **92,** 011103 (2004).
9. Ohnishi, N., Kotake, K. & Yamada, S. Numerical analysis on standing accretion shock instability with neutrino heating in the supernova cores. *Astrophys. J.* **641,** 1018–1028 (2006).
10. Blondin, J. M. & Mezzacappa, A. Spherical accretion shock instability in the linear regime. *Astrophys. J.* **642,** 401–409 (2006).
11. Lattimer, J. M. & Prakash, M. Neutron star structure and the equation of state. *Astrophys. J.* **550,** 426–442 (2001).
12. Liebendoerfer, M. *et al.* Probing the gravitational well: No supernova explosion in spherical symmetry with general relativistic Boltzmann neutrino transport. *Phys. Rev. D* **63,** 103004–103017 (2001).
13. Whitham, G. B. *Linear and Nonlinear Waves* 289 (Wiley, New York, 1974).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a SciDAC grant from the US Department of Energy High Energy, Nuclear Physics, and Advanced Scientific Computing Research Programs. A.M. is supported at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy. The simulations presented here were performed at the Leadership Computing Facility at ORNL. We thank the National Center for Computational Sciences at ORNL for their resources and support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.M.B. (john_blondin@ncsu.edu).

LETTERS

The lakes of Titan

E. R. Stofan[1,2], C. Elachi[3], J. I. Lunine[4], R. D. Lorenz[5], B. Stiles[3], K. L. Mitchell[3], S. Ostro[3], L. Soderblom[6], C. Wood[7], H. Zebker[8], S. Wall[3], M. Janssen[3], R. Kirk[6], R. Lopes[3], F. Paganelli[3], J. Radebaugh[4], L. Wye[8], Y. Anderson[3], M. Allison[9], R. Boehmer[3], P. Callahan[3], P. Encrenaz[10], E. Flamini[11], G. Francescetti[12], Y. Gim[3], G. Hamilton[3], S. Hensley[3], W. T. K. Johnson[3], K. Kelleher[3], D. Muhleman[13], P. Paillou[14], G. Picardi[15], F. Posa[16], L. Roth[3], R. Seu[15], S. Shaffer[3], S. Vetrella[12] & R. West[3]

The surface of Saturn's haze-shrouded moon Titan has long been proposed to have oceans or lakes, on the basis of the stability of liquid methane at the surface[1,2]. Initial visible[3] and radar[4,5] imaging failed to find any evidence of an ocean, although abundant evidence was found that flowing liquids have existed on the surface[5,6]. Here we provide definitive evidence for the presence of lakes on the surface of Titan, obtained during the Cassini Radar flyby of Titan on 22 July 2006 (T_{16}). The radar imaging polewards of 70° north shows more than 75 circular to irregular radar-dark patches, in a region where liquid methane and ethane are expected to be abundant and stable on the surface[2,7]. The radar-dark patches are interpreted as lakes on the basis of their very low radar reflectivity and morphological similarities to lakes, including associated channels and location in topographic depressions. Some of the lakes do not completely fill the depressions in which they lie, and apparently dry depressions are present. We interpret this to indicate that lakes are present in a number of states, including partly dry and liquid-filled. These northern-hemisphere lakes constitute the strongest evidence yet that a condensable-liquid hydrological cycle is active in Titan's surface and atmosphere, in which the lakes are filled through rainfall and/or intersection with the subsurface 'liquid methane' table.

Liquid methane is a thermodynamically allowed phase anywhere on the surface of Titan today. However, at all except the highest latitudes, the methane relative humidity (amount of methane relative to the saturated value) is less than 100%, and so standing bodies of methane must evaporate into the atmosphere. There is no methane ocean in contact with the atmosphere[8], and the timescale for saturating the atmosphere by evaporation of methane from the surface (about 10^3 years (ref. 9)) is much longer than the seasonal cycle of just under 30 years. Hence methane precipitation near the poles should dominate the 'hydrology' of methane on Titan[10]. Thus, lakes will be stable from the poles down to a latitude determined by the abundance of methane in the surface–atmosphere system and by the possible intersection of such surface methane fluids with putative subterranean 'methanifers', analogous to terrestrial aquifers.

An additional factor in establishing and stabilizing the presence of lakes at high latitude is the preferential deposition of ethane in polar regions (see, for example, ref. 11). This in turn may be controlled by the availability of cloud condensation nuclei in the stratosphere enhanced by the sedimentation of heavier organics such as C_4N_2 (ref. 12) in the seasonal polar hood. This feature, imaged nearly a Titan year ago by Voyager in the north and now also seen to be present (see, for example, ref. 13), forms during polar winter. Observations and modelling suggest that high-latitude clouds poleward of 75° south are methane but include an ethane mist[9]. Ethane is fairly involatile at Titan surface conditions, and hence if present would form a permanent component to lakes. With the observations currently available, the condensed-phase surface methane-to-ethane ratio cannot be constrained. Even if the as yet unmeasured surface temperatures above 75° north latitude are 3–5 K below the equatorial 93.6 K (ref. 14), as suggested by Voyager observations[11], dissolved nitrogen in binary methane–nitrogen lakes will depress the freezing point sufficiently, and ternary methane–ethane–nitrogen lakes will have strongly depressed freezing points.

The Cassini Titan Radar Mapper[4,15] (K_u-band wavelength 2.17 cm) instrument had its sixth radar pass of Titan (T_{16}) on 22 July 2006 (UTC). The synthetic aperture radar (SAR) arc-shaped imaging swath extends from mid-northern latitudes to near the north pole and back, and is 6,130 km long with spatial resolutions of 300–1,200 m (Fig. 1, and Supplementary Fig. 1). Incidence angles across the swath vary from 15° to 35°. The portion of the swath that extended from about 70° to 83° north contained more than 75 radar dark patches, from 3 km to more than 70 km across.

The dark patches contrast with the surrounding terrain, which has a mottled appearance similar to that of other 'plains' regions on Titan[4,5,16]. The backscatter of some of the dark patches is extremely low. Several appear at the noise floor of the data (about -25 dB σ_0), with much lower reflectivity than previously imaged areas on Titan, including the radar-dark (about -13 dB σ_0) sand dunes observed near Titan's equator. For the darkest patch we have observed so far, the normalized radar cross-section (σ_0) value is less than about -26 dB and could be zero, because the measured signal is at the system noise level (Fig. 2). We are unable to ascertain that any signal at all has been reflected from this feature.

The radar backscatter of the dark patches at Cassini SAR imaging incidence angles is consistent with that expected from a very smooth surface of any kind (for example liquid, rock, ice or organics) or even simply a non-reflecting, absorbing surface (for example a low-density surface smoothly matched into a non-scattering absorber such as fluffy soot or dirty snow overlying a uniform and electrically absorbing substrate). Radiometric brightness temperatures are obtained along with the SAR swaths, although the spatial resolution is limited to the footprints of the respective radar beams (that is, more than

[1]Proxemy Research, Rectortown, Virginia 20140, USA. [2]Department of Earth Sciences, University College London, London WC1E 6BT, UK. [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. [4]Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. [5]Space Department, Johns Hopkins University Applied Physics Lab, Laurel, Maryland 20723-6099, USA. [6]US Geological Survey, Flagstaff, Arizona 86001, USA. [7]Wheeling Jesuit University and Planetary Science Institute, Tucson, Arizona 85719, USA. [8]Stanford University, Stanford, California 94305, USA. [9]Goddard Institute for Space Studies, National Aeronautics and Space Administration New York, New York 10025, USA. [10]Observatoire de Paris, 92195 Meudon, France. [11]Alenia Aerospazio, 00131 Rome, Italy. [12]Facoltá di Ingegneria, 80125 Naples, Italy. [13]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [14]Observatoire Aquitain des Sciences de l'Univers UMR 5804, 33270 Floirac, France. [15]Universitá La Sapienza, 00184 Rome, Italy. [16]Dipartimento Interateneo di Fisica, Politecnico di Bari, 70126 Bari, Italy.

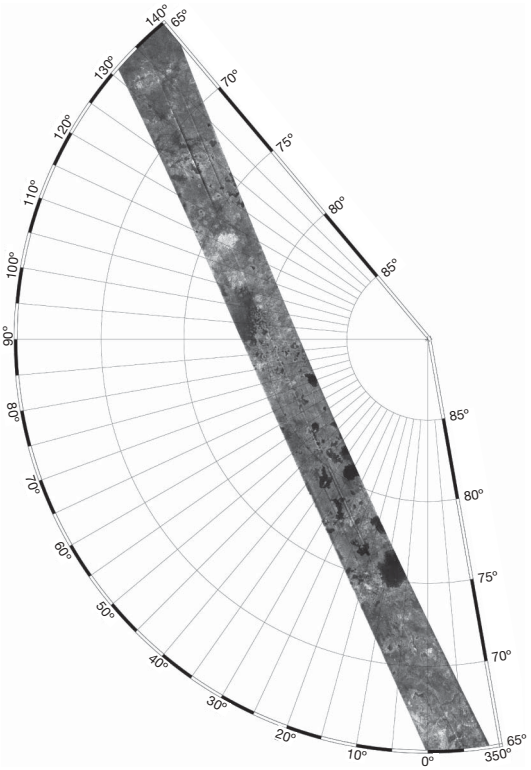


Figure 1 | Northern portion of the T_{16} swath. This portion of the T_{16} swath contains the dark patches interpreted as lakes. The image swath is shown in a polar projection; for scale, 1° of latitude is about 45 km. A high-resolution version of this image is given in Supplementary Information.

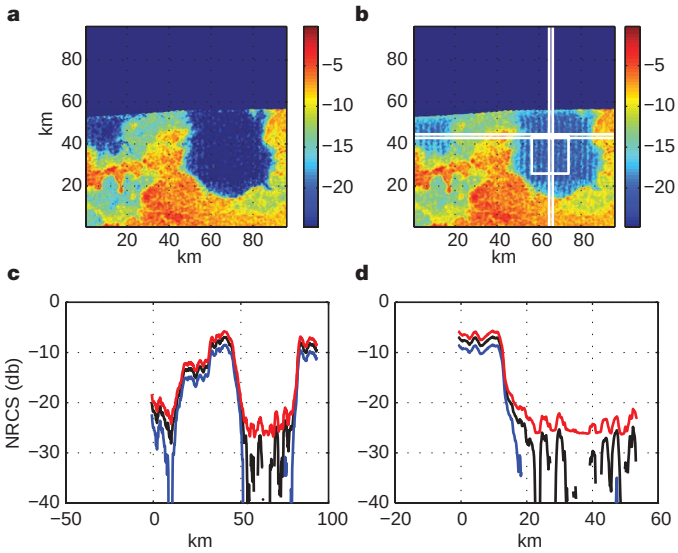


Figure 2 | Radar return from the darkest observed lake. **a**, Pseudo-colour image of observed normalized radar cross-section (NRCS) over lake at 80.5° north, 50° west. Corrections have been applied to remove known systematic biases in return resulting from thermal noise. NRCS measurement produced by thermal noise alone is referred to as the noise floor. Noise floor varies throughout the image due primarily to shape of antenna gain pattern. Dark blue region at top is outside radar observation. **b**, The same image without noise correction. White vertical and horizontal lines depict cuts used to produce **c** and **d**. To reduce the effect of random variance, we averaged NRCS over a 17 m × 17 km region (white rectangle). **c**, Horizontal cut through the noise-corrected NRCS image (black) with a 2σ lower bound (blue) and upper bound (red). In **c** and **d**, each point on the solid black curve is the average NRCS over a 3.6 km × 3.6 km area. Error bars include the effect of relative calibration errors due to spacecraft attitude knowledge, errors in noise subtraction process, and random variance. Cassini radar NRCS estimates have an unknown absolute bias estimated to be ±3 dB. Gaps in the NRCS curve and the lower bound curve are regions in which the values are zero or negative and thus cannot be represented in dB. **d**, Vertical cut through the noise-corrected NRCS image. There is a large contrast (15 dB) in NRCS across lake boundaries.

6 km). Nevertheless, several dark patches were observed that were resolved or partly resolved by at least one of the beams. The temperature contrast between these dark patches and the surrounding terrain is consistent with a flat surface of low dielectric constant in an ice terrain, where the emissivity of the ice is probably decreased somewhat by volume scattering. From microwave reflectivity and emissivity considerations, then, the dark patches are probably smooth surfaces of a low-dielectric material such as liquid methane.

At several of the dark patches, radar-dark sinuous features lead into the dark patch (Fig. 3a). These features resemble channels elsewhere on Titan interpreted to be fluvial in origin[4,5,17]. Two of the dark patches are connected by a narrow, radar-dark channel (Fig. 3b). However, even the dark patches with channels do not have extensive associated channel networks, suggesting that the channels are not the sole source of dark material infilling the patches.

Fifteen dark patches are in relatively steep-sided, rimmed, circular depressions, in contrast with other dark patches in this region that exhibit no such topography at this scale. These depressions seem to have existed in this form before being filled with dark material, and do not show clear evidence of erosion. The dark patches in depressions resemble terrestrial lakes confined within impact basins (for example Clearwater Lakes, Canada), volcanic calderas (for example Crater Lake, Oregon, USA) or karst dolines or sinkholes. The nested nature and limited size range of the depressions make an impact origin unlikely. A volcanic origin for the depressions is possible, given their morphology and our previous identification of cryovolcanic calderas on Titan[4,18]. Although liquid methane is thought to be unable to dissolve water ice[19], it is also possible that dissolution of a mixed organics and ice substrate has produced topographic depressions that can hold liquid.

Only two hypotheses are consistent with the radiometric and morphological characteristics of the dark patches: either we are observing liquid-filled lakes on Titan today, or depressions and channels formed in the past have now been infilled by a very low-density deposit that is darker than any observed elsewhere on Titan. The absence of any aeolian features in this area makes low-density, porous, unconsolidated sediments unlikely. This, combined with the morphologic characteristics of the dark patches, leads us to conclude that the dark patches are lakes containing liquid hydrocarbons.

Several types of characteristic margins or shorelines of the lakes are observed (Figs 1 and 3). As described above, some have steep margins and very distinct edges, suggesting confinement by a topographic rim. These lakes are more consistent with seepage or groundwater drainage lakes, with the lake intersecting the subsurface liquid-methane table. Other lakes have diffuse, more scalloped edges, with a gradual decrease in backscatter towards the centre of the lake. These lakes are more likely to have associated channels, and thus may be either drainage lakes or groundwater drainage lakes. Other lakes are more sinuous, or have sinuous extensions, similar in appearance to terrestrial flooded river valleys.

Several of the depressions seem to be filled with liquid, whereas others are only partly filled (Fig. 3c). These partly filled depressions may never have filled fully, or may have partly evaporated at some point in the past. Other features in the swath have margins similar to other lakes but a surface backscatter similar to the surrounding terrain (Fig. 3d). We interpret these to be possibly depressions that are now devoid of liquid. These lakes in possible varying states of fill suggest that the lakes in this region of Titan might be ephemeral, on some unknown timescale.

Although the lakes have generally very radar-dark surfaces, some have areas of increased radar brightness most commonly near their edges. The proximity to the edges suggests that the enhanced

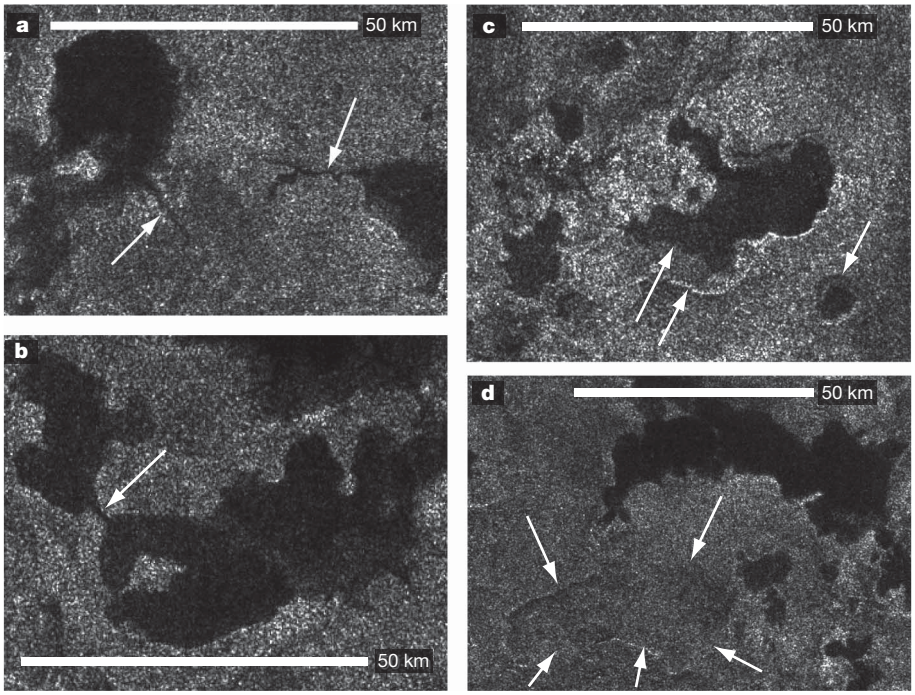


Figure 3 | Examples of lakes from the T_{16} swath. a, Sinuous radar-dark channels can be seen leading into two lakes. **b**, A pair of irregularly shaped lakes connected by a radar-dark channel. **c**, Radar-dark material in the lake to the left lies inside an apparent topographic edge (arrows), indicating that it might once have been at a higher level. The circular lake to the right seems to be a nested depression, with dark material filling the inner depression. **d**, A feature with a shape similar to lakes (arrows), but with backscatter similar to surrounding plains. On the basis of similarly shaped lakes and possibly partly filled features, we interpret this feature to be a dry lakebed.

brightness might be due to a reflection from the lake bottom, where it is sufficiently shallow that the bottom echo is not completely attenuated. K_u-band energy may penetrate many tens of metres through pure hydrocarbon liquids such as ethane or methane. Tidal winds at these high latitudes are predicted[20] to be less than 0.5 m s^{-1}, an order of magnitude lower than that needed to form capillary waves in liquid hydrocarbons in wind-tunnel tests on Earth[21]. The higher air density on Titan may facilitate wave growth somewhat, and over long fetches small gravity waves may form (amplitude about 5 cm; wavelength much less than 1 m) that could potentially be detected by the radar. We note that it is also possible that bright patches near the lake edges could be small 'islets' protruding through the surface. Floating 'icebergs' are unlikely because most materials would not float in liquid hydrocarbons.

Our inference that the northern-hemisphere lakes discovered by Cassini radar are at least partly liquid methane is consistent with various other considerations. If such lakes cover at least 0.2–4% of Titan's surface (depending on the amount of relatively involatile but highly soluble ethane contained within them), they will buffer the atmospheric methane's relative humidity at its observed value[7], removing the requirement for a putative steady drizzle at the equator[22]. If the abundance of lakes seen in the T_{16} data are typical of their coverage poleward of about 70° in both hemispheres, then the fraction of Titan's surface covered by lakes is within this range. More recent polar radar data from Cassini support this assertion.

Titan's northern hemisphere lakes probably participate in a methanological cycle that has multiple timescales. As Titan's seasons progress over the 29-year cycle of Saturn's orbit around the Sun, lakes in the winter hemisphere should expand by steady methane precipitation while summer hemisphere lakes shrink or dry up entirely. More speculatively and possibly less frequently (about 10^3 years for any given location[9]), mid-latitude and equatorial regions could experience a progressive growth in methane humidity leading to much more violent methane thunderstorms[23], carving the erosional patterns seen in the Huygens probe images[6] and other radar swaths[24]. On timescales of tens of millions of years[25], atmospheric photochemistry decreases the total methane—atmospheric, lacustrine and methaniferic—available to the system, causing lakes at progressively higher latitudes to be dry all year round. The abrupt and striking transition southwards of about 70° north latitude from the very dark lakes to features similarly shaped but bearing no contrast with the surroundings might be a consequence of the progressive depletion of methane from the surface–atmosphere system. On unknown but possibly longer timescales, thermal evolution of Titan's interior may drive additional methane into the surface–atmosphere system[26] by means of cryovolcanic events, geysering or the impact-generated release of methane stored below the surface. This last postulated resupply of methane from the interior is most speculative and is not directly implied by the presence of the lakes, although possible volcanic constructs seen by radar[4] and the visible and near-infrared mapping spectrometer[27] hint at the long-term role of volcanism and outgassing.

Future radar observations will determine the origin of lake surface textures (for example winds or lake bottom effects) and will also constrain liquid dielectric properties, and possibly depth and shoreline characteristics through diversity in the viewing geometry. SAR imagery of the lakes near the end of a proposed extended mission, in 2009 or 2010, would provide a sufficient time base on which to detect seasonally driven changes in lake extent, predicted to occur[7] if the lakes are not connected to a much larger subsurface methanifer. Future missions to the surface will be required for a full understanding of the lakes of Titan, in particular how they formed, their detailed composition and their interaction with their shorelines.

Received 2 September; accepted 9 November 2006.

1. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
2. Lorenz, R. D., Kraal, E., Asphaug, E. & Thomson, R. The seas of Titan. *Eos* **84**, 125–132 (2003).
3. Porco, C. C. *et al.* Imaging of Titan from the Cassini spacecraft. *Nature* **434**, 159–168 (2005).
4. Elachi, C. *et al.* Cassini radar views the surface of Titan. *Science* **308**, 970–974 (2005).

5. Elachi, C. *et al.* Titan radar mapper observations from Cassini's T_3 flyby. *Nature* **441,** 709–713 (2006).
6. Tomasko, M. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* **438,** 765–778 (2005).
7. Mitri, G., Showman, A. P., Lunine, J. I. & Lorenz, R. D. Hydrocarbon lakes on Titan. *Icarus* (in the press).
8. West, R. A., Brown, M. E., Salinas, S. V., Bouchet, A. H. & Roe, H. G. No oceans on Titan from the absence of a near-Infrared specular reflection. *Nature* **436,** 670–672 (2005).
9. Lorenz, R. D., Griffith, C. A., Lunine, J. I., McKay, C. P. & Renno, N. O. Convective plumes and the scarcity of Titan's clouds. *Geophys. Res. Lett.* **32,** L01201 (2005).
10. Rannou, P., Montmessin, F., Hourdin, F. & Lebonnois, S. The latitudinal distribution of clouds on Titan. *Science* **311,** 201–205 (2006).
11. Samuelson, R. E., Nitya, N. R. & Borysow, A. Gaseous abundances and methane supersaturation in Titan's troposphere. *Planet. Space Sci.* **45,** 959–980 (1997).
12. Samuelson, R. E., May, L. A., Knuckles, M. A. & Khanna, R. J. C_4N_2 ice in Titan's north polar atmosphere. *Planet. Space Sci.* **45,** 941–948 (1997).
13. Lorenz, R. D., Lemmon, M. T. & Smith, P. H. Seasonal evolution of Titan's dark polar hood: midsummer disappearance observed by the Hubble Space Telescope. *Mon. Not. R. Astron. Soc.* **369,** 1683–1687 (2006).
14. Fulchignoni, M. *et al. In situ* measurements of the physical characteristics of Titan's environment. *Nature* **438,** 785–791 (2005).
15. Elachi, C., Im, E., Roth, L. E. & Werner, C. L. Cassini Titan radar mapper. *Proc. IEEE* **79,** 867–880 (1991).
16. Stofan, E. R. *et al.* Mapping of Titan: Results from the first Titan radar passes. *Icarus* (in the press).
17. Lorenz, R. D. *et al.* Fluvial channels on Titan: Meteorological paradigm and Cassini RADAR observations. *Planet. Space Sci.* (submitted).
18. Lopes, R. M. C. *et al.* Cryovolcanic features on Titan's surface as revealed by the Cassini Titan radar mapper. *Icarus* (in the press).
19. Lorenz, R. D. & Lunine, J. I. Erosion on Titan: Past and present. *Icarus* **122,** 79–91 (1996).
20. Tokano, T. & Neubauer, F. M. Wind-induced seasonal angular momentum exchange at Titan's surface and its influence on Titan's length-of-day. *Geophys. Res. Lett.* **32,** L24203 (2005).
21. Lorenz, R. D., Kraal, E., Eddlemon, E., Cheney, J. & Greeley, R. Sea-surface wave growth under extraterrestrial atmospheres—preliminary wind tunnel experiments with application to Mars and Titan. *Icarus* **175,** 556–560 (2005).
22. Tokano, T. *et al.* Methane drizzle on Titan. *Nature* **442,** 432–435 (2006).
23. Hueso, R. & Sánchez-Lavega, A. Methane storms on Saturn's moon Titan. *Nature* **442,** 428–431 (2006).
24. Lunine, J. I. *et al.* Cassini radar's third and fourth looks at Titan. *Icarus* (submitted).
25. Yung, Y. L., Allen, M. A. & Pinto, J. P. Photochemistry of the atmosphere of Titan: comparison between model and observations. *Astrophys. J. Suppl. Ser.* **55,** 465–506 (1984).
26. Tobie, G., Lunine, J. I. & Sotin, C. Episodic outgassing as the origin of atmospheric methane on Titan. *Nature* **440,** 61–64 (2006).
27. Sotin, C. *et al.* Release of volatiles from a possible cryovolcano from near-infrared imaging of Titan. *Nature* **435,** 786–789 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We gratefully acknowledge the long years of work by the entire Cassini team that allowed these data of Titan to be obtained. The Cassini Project is a joint endeavour of NASA, ESA and ASI, managed by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.R.S. (estofan@proxemy.com).

LETTERS

High-speed linear optics quantum computing using active feed-forward

Robert Prevedel[1], Philip Walther[1,2], Felix Tiefenbacher[1,3], Pascal Böhi[1]†, Rainer Kaltenbaek[1], Thomas Jennewein[3] & Anton Zeilinger[1,3]

As information carriers in quantum computing[1], photonic qubits have the advantage of undergoing negligible decoherence. However, the absence of any significant photon–photon interaction is problematic for the realization of non-trivial two-qubit gates. One solution is to introduce an effective nonlinearity by measurements resulting in probabilistic gate operations[2,3]. In one-way quantum computation[4–8], the random quantum measurement error can be overcome by applying a feed-forward technique, such that the future measurement basis depends on earlier measurement results. This technique is crucial for achieving deterministic quantum computation once a cluster state (the highly entangled multiparticle state on which one-way quantum computation is based) is prepared. Here we realize a concatenated scheme of measurement and active feed-forward in a one-way quantum computing experiment. We demonstrate that, for a perfect cluster state and no photon loss, our quantum computation scheme would operate with good fidelity and that our feed-forward components function with very high speed and low error for detected photons. With present technology, the individual computational step (in our case the individual feed-forward cycle) can be operated in less than 150 ns using electro-optical modulators. This is an important result for the future development of one-way quantum computers, whose large-scale implementation will depend on advances in the production and detection of the required highly entangled cluster states.

One-way quantum computation is based on highly entangled multiparticle states, so-called cluster states, which are a resource for universal quantum computing. On these cluster states, single-qubit measurements alone are sufficient to implement universal quantum computation. Different algorithms only require a different 'pattern' of single-qubit operations on a sufficiently large cluster state; as explained in ref. 5, ''the cluster states are one-way quantum computers and the measurements form the program.'' In contrast to the standard linear optics architecture, which relies on multiparticle gates, the cluster state computation is performed by consecutive single-qubit measurements where the choice of the future measurement basis is dependent on the outcome of preceding measurements. Active feed-forward of the classical measurement results renders one-way quantum computation deterministic—that is, given a perfect cluster state and exact measurements, the processing of encoded information on physical qubits is accomplished without error. The one-way quantum computer model that we employ is currently the only one that promises deterministic photonic quantum computation (through feed-forward). Standard optical schemes[2] achieve this only in the asymptotic regime of numerous gates and/or photons. Nevertheless, we note that feed-forward control based on measurements made on ancillary qubits is also essential for error correction in the standard network approach.

Recently, the working principles of one-way quantum computing have been demonstrated using cluster states encoded into the polarization states of photons[9–11]. However, all experiments so far have been performed using fixed polarizer settings, thus making the computation probabilistic (that is, not scalable) and wasting precious resources on the way. In this Letter we demonstrate feed-forward linear optics quantum computation on a four-qubit cluster state. The cluster state creation is based on a post-selection technique developed in ref. 9, and the feed-forward stages are realized by employing fibre delays and fast active switches for selecting the appropriate measurement basis and correcting introduced Pauli errors. Earlier proof-of-principle demonstrations[12–14] of feed-forward control were limited to two photons and one feed-forward step only. However, to demonstrate feed-forward quantum computing, more photons and thus several consecutive feed-forward steps are required. It is particularly important to realize a situation where a later measurement depends on an earlier measurement and its fed-forward result. Dealing with the complex situation of a four-qubit cluster state and three electro-optical modulators (EOMs), we demonstrate 'error-free' single-qubit and two-qubit gate operations as well as Grover's search algorithm[15].

Given a cluster state, two basic types of single-particle measurements suffice to operate the one-way quantum computer. Measurements in the computational basis $\left\{|0\rangle_j,|1\rangle_j\right\}$ have the effect of disentangling, that is, removing the physical qubit j from the cluster, thus leaving a smaller cluster state. The measurements that perform the actual quantum information processing, however, are made in the basis $B_j(\alpha)=\left\{|\alpha_+\rangle_j,|\alpha_-\rangle_j\right\}$, where $|\alpha_\pm\rangle_j=\left(e^{i\alpha/2}|0\rangle_j\pm e^{-i\alpha/2}|1\rangle_j\right)/\sqrt{2}$ with $\alpha\in\{0,2\pi\}$. The choice of measurement basis determines the single-qubit rotation, $R_z(\alpha)=\exp(-i\alpha\sigma_z/2)$, followed by a Hadamard operation, $H=(\sigma_x+\sigma_z)/\sqrt{2}$, of encoded qubits in the cluster[9] ($\sigma_x,\sigma_y,\sigma_z$ being the Pauli matrices). Any quantum logic operation can be carried out by an appropriate choice of $B_j(\alpha)$ on a sufficiently large cluster state. We define the outcome s_j of a measurement on the physical qubit j to be '0' if the measurement outcome is $|\alpha_+\rangle_j$ and '1' if the outcome is $|\alpha_-\rangle_j$. Whenever the outcome is '0', the computation proceeds without error, whereas for the case where the outcome is '1', a well-defined Pauli error is introduced. These known errors are compensated for by feed-forward such that the output controls future measurement settings.

In the present work, we create a cluster state of the form:

$$|\Phi_{\text{cluster}}\rangle=\frac{1}{2}\left(|0\rangle_1|0\rangle_2|0\rangle_3|0\rangle_4+|0\rangle_1|0\rangle_2|1\rangle_3|1\rangle_4+|1\rangle_1|1\rangle_2|0\rangle_3|0\rangle_4-|1\rangle_1|1\rangle_2|1\rangle_3|1\rangle_4\right) \quad (1)$$

[1]Institute for Experimental Physics, University of Vienna, Boltzmanngasse 5, A-1090 Vienna, Austria. [2]Physics Department, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria. †Present address: Max-Planck-Institut für Quantenoptik und Sektion Physik der Ludwig-Maximilians-Universität, Schellingstr. 4, 80799 München, Germany.

where $|0\rangle$ and $|1\rangle$, in the actual experiment, denote horizontal and vertical polarization, respectively (the subscript labels the photon). The state of equation (1) is equivalent to the four-qubit linear cluster $|\Phi_{\text{lin4}}\rangle$ and to the horseshoe cluster[9] $|\Phi_{\subset 4}\rangle$ under the local unitary operation $H_1 \otimes I_2 \otimes I_3 \otimes H_4$ on the physical cluster state (H is the Hadamard and I is the identity operation). In the experiment, the cluster state is known to have been prepared when all four photons are detected. This ensures that photon loss and photodetector inefficiency do not affect the experimental results. The state creation is verified by over-complete state tomography in which the density matrix of the cluster state is reconstructed from a set of 1,296 local measurements using a maximum-likelihood technique[16,17] and all combinations of mutually unbiased basis sets for individual qubits, that is, $\{|0\rangle,|1\rangle;|+\rangle,|-\rangle;|R\rangle,|L\rangle\}$, where $|+/-\rangle = (|0\rangle \pm |1\rangle)/\sqrt{2}$ denote $\pm 45^\circ$ polarization and $|L/R\rangle = (|0\rangle \pm i|1\rangle)/\sqrt{2}$ stands for right and left circular polarization. Each of these measurements took 500 s. The experimentally obtained density matrix, ρ, has a fidelity of $F = \langle \Phi_{\text{cluster}}|\rho|\Phi_{\text{cluster}}\rangle = (0.62 \pm 0.01)$ with the ideal four-qubit cluster, $|\Phi_{\text{cluster}}\rangle$, which despite all EOMs and fibre-coupled outputs is sufficiently above the threshold for entanglement[18] of 0.5.

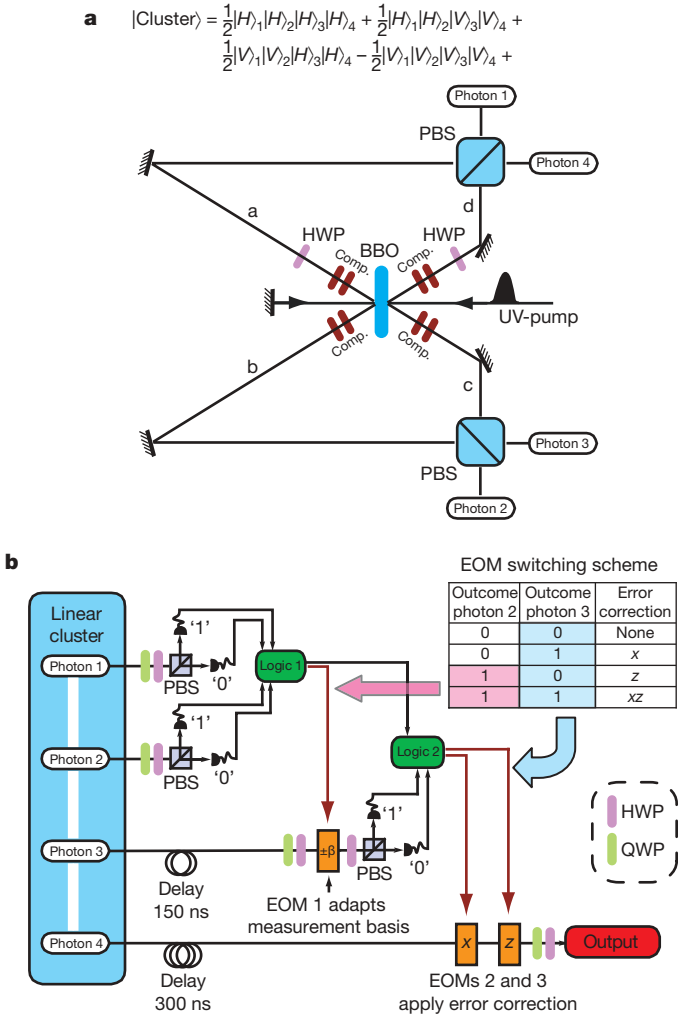


Figure 1 | Schematic drawing of the experimental set-up. Interferometric cluster-state preparation is shown in **a**, and its extension to achieve active feed-forward of the one-way quantum computation is depicted in **b**. An ultraviolet (UV) laser pulse passes twice through a nonlinear crystal to produce polarization-entangled photon pairs in both the forward and backward direction. Compensators (Comp.) are half-wave plates (HWP) and BBO crystals used to counter walk-off effects in the down-conversion crystal. They are aligned such that $|\Phi^-\rangle$ and $|\Phi^+\rangle$ states are emitted in the forward and backward direction, respectively. Taking into account the possibility of double-pair emission and the action of the polarizing beam splitters (PBSs), the four amplitudes of the linear cluster state can be generated with an additional HWP in mode a. Once this is achieved, the computation proceeds by consecutive polarization measurements on photons 1–4. Dependent on the outcomes of photons 1–3, three fast electro-optical modulators (EOMs) are employed to implement the active feed-forward. One EOM adapts the measurement basis of photon 3, while two EOMs, aligned for σ_x and σ_z operation, apply the error correction on output photon 4. Two single-mode fibres, 30 m and 60 m long, serve to locally delay the photons during the detection stage, logics operation and switching/charging process of the EOMs. The polarization of the photons is measured by a PBS preceded by a HWP and a quarter-wave plate (QWP) in each mode.

Using present technologies and customized fast EOMs, we were able to realize high-fidelity (>99% for detected photons) fast active switching with feed-forward times of less than 150 ns (Methods). The gate operation that can therefore be achieved is about three orders of magnitude faster than comparable physical realizations of quantum computers[19–21]. The measurement device for an arbitrary basis consists of a quarter-wave and a half-wave plate followed by a polarizing beam-splitter (PBS), which transmits horizontally polarized light ('0') and reflects vertically polarized light ('1'). While qubit 1 and qubit 2 are measured without any delay, qubit 3 and qubit 4 are delayed in optical single-mode fibres with lengths of 30 m (150 ns) and 60 m (300 ns), respectively. The active switching itself is achieved via Pockels cells; one for qubit 3 to adapt the measurement basis, that is, from $B_3(\beta)$ to $B_3(-\beta)$, and two in the channel of output-qubit 4 to correct introduced Pauli errors, σ_x and σ_z (Fig. 1).

As an example, consider the general case of a three-qubit linear cluster state $|\Phi_{\text{lin3}}\rangle$, such as that depicted in Fig. 2a. This state can be obtained from our four-qubit cluster by removing qubit 1, that is, measuring this qubit in the computational basis for the linear cluster, $\{|+\rangle_1,|-\rangle_1\}$. Consecutive measurements in bases $B_2(\alpha)$ and $B_3(\beta)$ on the physical qubits 2 and 3 implement an arbitrary single-qubit rotation of the encoded input qubit $|\Psi_{\text{in}}\rangle = |+\rangle_1$. These measurements rotate the encoded input qubit to the output state $|\Psi_{\text{out}}\rangle = \sigma_x^{s_3} HR_z((-1)^{s_2}\beta)\sigma_x^{s_2} HR_z(\alpha)|\Psi_{\text{in}}\rangle = \sigma_x^{s_3}\sigma_z^{s_2} R_x((-1)^{s_2}\beta)R_z(\alpha)|\Psi_{\text{in}}\rangle$, which is stored on qubit 4. The measurement outcome, $s_i = \{0,1\}$, on the physical qubit i, (1) determines the measurement basis for the succeeding qubit, and (2) indicates any introduced Pauli errors that have to be compensated for. In the specific case where the outcomes of the second and third qubit are $s_2 = s_3 = 0$, no error correction is required: $|\Psi_{\text{out}}\rangle = R_x(\beta)R_z(\alpha)|\Psi_{\text{in}}\rangle$. Whenever the outcome of the second qubit is $s_2 = 1$ ($s_3 = 0$), then the measurement basis of the third qubit has to be changed from $B_3(\beta)$ to $B_3(-\beta)$ and finalized by a Pauli error correction, σ_z, that is, $|\Psi_{\text{out}}\rangle = \sigma_z R_x(-\beta)R_z(\alpha)|\Psi_{\text{in}}\rangle$, to get the same output as if no error had occurred. Similar corrections are required in the cases when the third qubit's outcome is $s_3 = 1$ ($s_2 = 0$): $|\Psi_{\text{out}}\rangle = \sigma_x R_x(\beta)R_z(\alpha)|\Psi_{\text{in}}\rangle$ or, if an unwanted projection occurs to both qubits, $s_2 = s_3 = 1$, two Pauli errors, σ_z and σ_x, have to be compensated for on qubit 4, $|\Psi_{\text{out}}\rangle = \sigma_x\sigma_z R_x(-\beta)R_z(\alpha)|\Psi_{\text{in}}\rangle$. The same feed-forward techniques hold for two-dimensional cluster states (Fig. 3), which will be discussed later in the paper.

Examples of single-qubit rotations with feed-forward are shown in Fig. 2, together with the outcomes of the same computation in the case when no feed-forward is applied. In each case, the output of the single-qubit rotation is stored in qubit 4 and completely characterized by single-qubit tomography. Figure 2a shows a diagram of the implemented quantum algorithm; Fig. 2b shows the output of the computation $|\Psi_{\text{out}}\rangle = R_x\left(-\frac{\pi}{2}\right)R_z\left(-\frac{\pi}{2}\right)|\Psi_{\text{in}}\rangle = |+\rangle_4$, in the laboratory basis, with and without active feed-forward. We find an average fidelity of (0.84 ± 0.08) with the ideal state when active feed-forward is implemented. This is a considerable improvement over the case of no feed-forward, which produces the target state with an average fidelity of only (0.55 ± 0.06). In order to prove universal quantum computing, we need to demonstrate single-qubit rotations outside the Clifford group[22]. This special example is shown in Fig. 2c, where we perform polarization projections in the basis $\alpha = \frac{\pi}{4}$

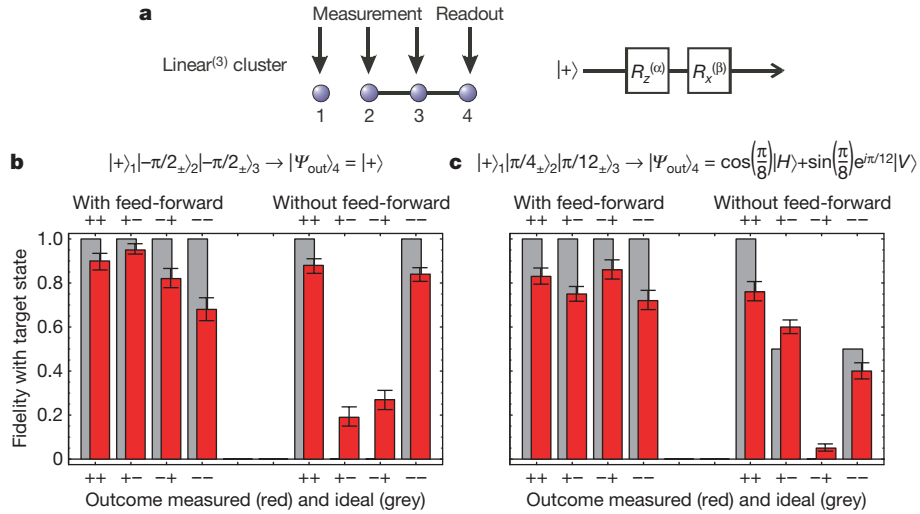


Figure 2 | Active feed-forward of two different single-qubit rotations. **a**, The linear three-qubit cluster state (obtained from our four-qubit cluster state) and the quantum circuit it implements. The operation $R_x(\alpha) = \exp(-i\alpha\sigma_x/2)$ can be implemented through the matrix identity $R_x(\alpha) = HR_z(\alpha)H$. **b**, **c**, Fidelity of the output state with the desired state in the case of active feed-forward and without feed-forward of measurement results. Both the experimentally measured fidelities (red bars) and the expected, ideal fidelities (grey bars) are given. It is immediately clear that, with feed-forward, the computation theoretically always produces the desired outcome with certainty, even if measurement outcomes in the $|\alpha\rangle_2$, $|\beta\rangle_3$ basis deviate from the desired $s_2 = s_3 = 0$ event ('++'). In **b**, α and β were both set to $-\frac{\pi}{2}$, resulting in the output state $|\Psi_{out}\rangle_4 = |+\rangle$, while in **c**, the measurement angles were set to $\alpha = \frac{\pi}{4}$ and $\beta = \frac{\pi}{12}$. Averaged over all possible measurement outcomes, the overlap of the measured one-qubit density matrix with the ideal state with feed-forward is (0.84 ± 0.08) in **b** and (0.79 ± 0.07) in **c**, respectively. Without feed-forward, theory predicts an average fidelity of 0.5. In the experiment, we find (0.55 ± 0.06) and (0.45 ± 0.05), for **b** and **c**, respectively. Error bars indicate s.d.

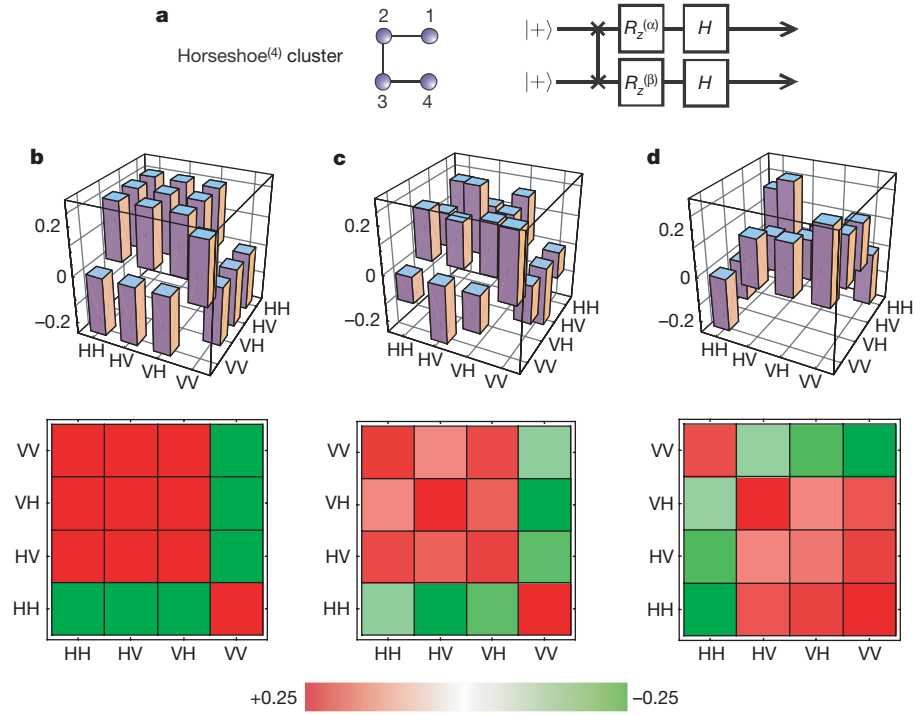


Figure 3 | Feed-forward of a two-qubit operation. We perform the operation $|+\rangle_{1E}|+\rangle_{2E} \rightarrow \frac{1}{\sqrt{2}}(|H\rangle_1|+\rangle_4 + |V\rangle_1|-\rangle_4)$ with single-qubit measurements in $B_2(0)$ and $B_3(0)$ carried out on photons 2 and 3 on the horseshoe cluster state $|\Phi_{\subset 4}\rangle$. **a**, The algorithm implemented by the horseshoe cluster. **b**, The ideal, expected density matrix, with the real part of the density matrix shown as a bar chart (upper figure) and as a coloured grid plot (lower figure). The imaginary components of the density matrices are zero in theory and negligible in the experiment. **c**, In the case where photon 2 and 3's outcome was $s_2 = s_3 = 1$ instead of the desired '00' event, the logical feed-forward relation has been applied by relabelling the analyser output ports. Fidelity and measures of entanglement of the reconstructed state can be inferred from the main text. In **d**, we show the output of the same quantum computation when no feed-forward is applied. The experimental density matrix in this case differs remarkably from the ideal one, which is reflected in the low state fidelity (see main text).

and $\beta = \frac{\pi}{12}$, which results in the more complex computation $|\Psi_{\text{out}}\rangle_4 = R_x\left(\frac{\pi}{4}\right)R_z\left(\frac{\pi}{12}\right)|\Psi_{\text{in}}\rangle = \cos\left(\frac{\pi}{8}\right)|H\rangle + \sin\left(\frac{\pi}{8}\right)e^{i\pi/12}|V\rangle$ in the error-free case. Here we find an average fidelity of (0.79 ± 0.07) with active feed-forward, but only (0.45 ± 0.05) without (Fig. 2). We find similar results for other measurement angles and hence other single-qubit rotations. The error bars of the above results were estimated by performing a 100 run Monte Carlo simulation of the whole state tomography analysis, with poissonian noise added to the count statistics in each run. Note that the reduced output state fidelity is mainly due to the non-ideal cluster state preparation and not due to erroneous switching of the EOMs, as these operate with very high precision.

It is a specific strength of the cluster-state computation that the adaptation of the measurement basis, $B_j(\alpha)$, caused by the measurement outcome of the preceding qubit, can be carried out without active switching when the eigenstates of the measurement basis are identical to the eigenstates of $\sigma_x\sigma_y$ or σ_z. In that case, logical feed-forward results in a reinterpretation of the measurement outcome. Outcome '0' would then correspond to the measurement outcome $|\alpha_-\rangle_j$ and '1' to the outcome $|\alpha_+\rangle_j$. We demonstrate this specific feature within the two-dimensional four-qubit cluster states, the horseshoe cluster, $|\Phi_{\subset 4}\rangle$, and the box cluster, $|\Phi_{\square 4}\rangle$, which we use to realize an entangling gate and an implementation of Grover's quantum search algorithm.

Universal quantum computing requires a universal set of one- and two-qubit operations such as the controlled-NOT (CNOT) or controlled-PHASE (CPhase) gates which can be realized using either horseshoe or box clusters. These gates can be implemented on our linear cluster by changing the order of measurements; for example, by measuring qubits 2 and 3 and thus transferring the two-qubit quantum state onto the remaining qubits 1 and 4. This quantum circuit can also be written as $|\Psi_{\text{out}}\rangle = \left(\sigma_x^{s_2}\otimes\sigma_x^{s_3}\right)(H_1\otimes H_2)[R_z(\alpha)\otimes R_z(\beta)]\text{CPhase}|\Psi_{\text{in}}\rangle$, where $|\Psi_{\text{in}}\rangle = |+\rangle_1|+\rangle_2$ is our encoded two-qubit input state. Note that the Pauli errors have to be compensated for in the case where $s_2 = 1$ and/or $s_3 = 1$. In principle, feed-forward relations in the case of two-qubit gates are more complex, as measurement errors in one 'rail' may influence the state of the qubit in another rail[23]. In particular, for polarization projections $|\alpha_+\rangle_2|\beta_+\rangle_3, |\alpha_+\rangle_2|\beta_-\rangle_3, |\alpha_-\rangle_2|\beta_+\rangle_3, |\alpha_-\rangle_2|\beta_-\rangle_3$ (that is, for measurement outcomes $s_2 = s_3 = 0$; $s_2 = 0$, $s_3 = 1$; $s_2 = 1$, $s_3 = 0$; $s_2 = s_3 = 1$), the operation $I_1 \otimes I_4$, $I_1 \otimes \sigma_{x4}$, $\sigma_{x1} \otimes I_4$, $\sigma_{x1} \otimes \sigma_{x4}$ has to be fed-forward to the output qubits 1 and 4, respectively. In Fig. 3, we explicitly show the case where both photons 2 and 3 are measured to be $s_2 = s_3 = 1$ instead of the desired '0' outcomes $s_2 = s_3 = 0$ in the bases $B_2(0)$ and $B_3(0)$. Those 'errors' rotate the input state to the maximally entangled output state $|\Psi_{\text{out}}\rangle_{1,4} = \frac{1}{\sqrt{2}}\left(|+\rangle_1|V\rangle_4 - |-\rangle_1|H\rangle_4\right)$. To obtain the desired state $|\Psi_{\text{out}}\rangle_{1,4} = \frac{1}{\sqrt{2}}\left(|H\rangle_1|+\rangle_4 + |V\rangle_1|-\rangle_4\right)$, the operation $\sigma_{x1} \otimes \sigma_{x4}$ has to be fed-forward on qubits 1 and 4. Density matrices of the ideal two-qubit output state and the experimentally reconstructed state are shown in Fig. 3, together with the measured output state obtained without feed-forward. We compute a state fidelity of (0.79 ± 0.04) for the overlap of our experimental fed-forward state with the ideal one. The tangle[24] of this output state is $\tau = (0.42 \pm 0.09)$, confirming the generation of entanglement between the output qubits as a result of the computation. Furthermore, our reconstructed density matrix implies a maximum CHSH Bell parameter[25] of $S = (2.40 \pm 0.09)$, which is more than four standard deviations above the $S = 2$ upper limit for local realistic theories. For comparison, if the feed-forward relation is not applied to this specific computation, the measured fidelity is only (0.09 ± 0.03), in agreement with the theoretical prediction of 0—no overlap with the desired state.

Quantum algorithms[15,26] are fascinating applications of quantum computers. Interestingly, Grover's quantum search algorithm[9,15,27] can be implemented on a four-qubit box cluster, such as the one depicted in Fig. 4a, with final readout measurements made in the basis $B_{1,4}(\pi)$ on physical qubits 1 and 4. Grover's algorithm promises a quadratic speed-up for unstructured search. It is worth mentioning that, for the case of four entries in an unsorted database[15], Grover's search will find the marked entry with certainty after a single iteration. The algorithm can be separated into two basic operations. First, a quantum device—the so-called 'oracle'—labels the correct element, which can be set by a proper choice of α and β; specifically, it tags one of the four computational basis states $|0\rangle|0\rangle$, $|0\rangle|1\rangle$, $|1\rangle|0\rangle$ and $|1\rangle|1\rangle$ by changing its sign, for example, $|0\rangle|0\rangle \rightarrow -|0\rangle|0\rangle$. Then, after an inversion-about-the-mean operation[9,15] the labelled element is found with certainty by the readout measurements. However, incorrect measurement outcomes at the 'oracle' (that is, at qubits 2 and 3) introduce Pauli errors, which effectively cause a wrong database element to be tagged. Feed-forward compensates for these errors such that the algorithm produces the right search result with certainty. In Fig. 4b, we show the experimental results of this quantum algorithm with and without feed-forward. The difference in performance is quite obvious: with feed-forward the right database element is identified with a probability of $(85\pm3)\%$, which compares favourably with the case when the feed-forward relation is not applied, which we find to be $(25\pm2)\%$, just as good as with an classical random search algorithm.

In summary, we have shown that in the absence of photon loss, a one-way quantum computer with active, concatenated feed-forward would operate with high fidelity. As in all current photonic quantum computation experiments, the input cluster state is produced

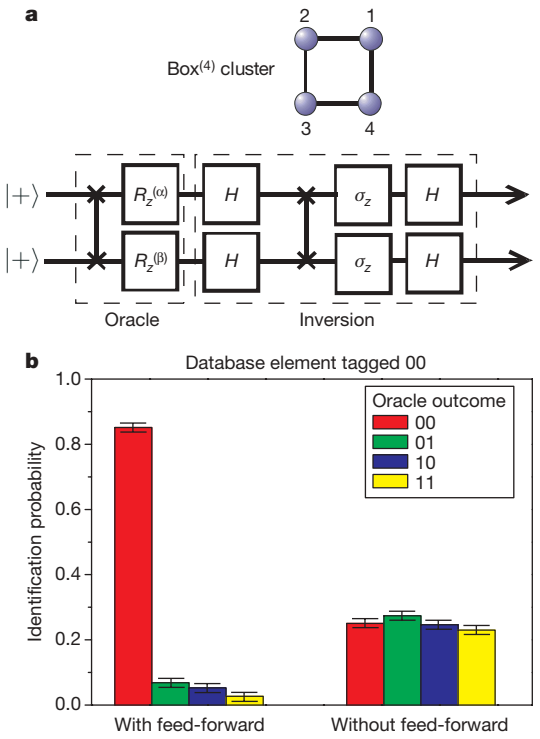


Figure 4 | Demonstration of Grover's search algorithm with feed-forward. Here we chose to tag the $|0\rangle|0\rangle$ entry. **a**, The algorithm consists of two distinct operations: The 'oracle' tags the unsorted database element by measuring physical qubits 2 and 3 in the bases $B_{2,3}(\pi)$, while the inversion process finds the desired database entry with certainty after a single query. Owing to intrinsic measurement randomness, however, it happens with equal probability that other database entries become tagged. Without feed-forward, on average this results in a balanced output of the algorithm, as can be seen from the experimental data in **b**. Applying the feed-forward procedure leads to an unambiguous search result, so that, on average, the algorithm finds the correct outcome with a probability of $(85 \pm 3)\%$. In the case without feed-forward, we find each possible result with equal probability of $(25 \pm 2)\%$. Error bars indicate s.d.

conditional on detecting all constituent photons. Because the efficiency of producing cluster states is low at present, this is not yet scalable. However, the technique is insensitive to photon loss due to absorption, reflection, fibre coupling and photodetector inefficiency. Thus our experiments show that except for photon loss, the feed-forward procedure operates with a quality and speed at present unmatched by other quantum computation methods. Conceptually, the most interesting result of our work is that it is indeed possible to build a deterministic quantum computer that has intrinsically random quantum measurements as its essential feature. Eventual large-scale implementations will need significant improvement of state preparation quality and photon detection efficiency, and reduction of photon losses. This will certainly be fostered by recent developments of highly efficient single-photon detection methods as well as 'on demand' single-photon sources. Given large and high-fidelity cluster states as well as low photon loss and significantly improved detectors, promising future applications of one-way quantum computers include important tasks such as the quantum Fourier transform[28,29], which is at the heart of Shor's factorizing algorithm.

METHODS

Experimental cluster state preparation. In our experiment, the cluster state is generated from photon pairs entangled in polarization and mode that originate from spontaneous parametric down-conversion. We employed the method of ref. 9 to generate the four-qubit cluster state, which is shown in Fig. 1 together with its extension to realize feed-forward quantum computation. The generation of the cluster state is dependent on simultaneous emission of four photons, that is, we only post-select those cases where each the four output modes a–d of the PBSs contain one photon (for more details, see the Methods section of ref. 9).

Contributions to feed-forward time. Quantum computation on a cluster state is performed by consecutive measurements on qubits 1–4. It is therefore necessary to locally delay photons 3 and 4 if active feed-forward of measurement results is desired. We find that the overall process of a single feed-forward step requires, on average, 145 ± 3 ns, where this value is composed of the following contributions: propagation time of photons 1 and 2 in single-mode fibres leading to detector (15 ns), delay of the single-photon detectors (35 ± 3 ns), processing time of the logic (7.5 ns), switching delay of the EOM driver (65 ns), rise time of the Pockels cell (5 ns), and miscellaneous coaxial cables employed in the set-up (17.5 ns). Two single-mode fibres, 30 m (150 ns) and 60 m (300 ns) long, serve to locally delay the photons during the detection stage, logic operations and the switching/charging process of the EOMs. We expect that the overall feed-forward time can be significantly reduced in optical fibres or waveguides where a smaller scale results in faster switching of the EOM driver[30].

Characterization of the feed-forward stage. For the active switching, we employ KD*P (potassium dideuterium phosphate) crystals with a measured transmission greater than 96%, a half-wave voltage of ~6.3 kV and a high switching contrast of approximately 500:1. The switching contrast is defined as the ratio of photons that are measured to obtain a well-defined polarization rotation due to the operation of the EOM divided by the photons that remain in the original state due to malfunctioning of the device. This was measured at the single photon level for various input polarizations employing time-correlated photons emitted by a down-conversion source, triggering on one photon and thereby rotating the polarization state of the other photon. From the high switching contrast of 500:1, one can infer that the total feed-forward accuracy of the three EOMs for detected photons is at least $(1 - 1/500)^3 > 0.99$. Other errors apart from photon loss, such as mode mismatch and unwanted phase shifts at the main PBSs, only result in non-ideally prepared input cluster states. However, the performance of the feed-forward stage is unaffected by these imperfections. In the present configuration, the custom-built EOM drivers can be operated up to 20 kHz; this is compatible with our trigger-rate requirement, which is set by our photon pair production rates (~2 kHz). During recharge cycles, the EOM drivers are 'disabled' for an effective dead time of 1.6 μs, which is short enough, considering our average two- and four-photon production rate (of the order of 2 kHz and 1 Hz, respectively). The overall detection efficiency of the experiment—bearing in mind the non-ideal collection of photons in single-mode fibres (~20%), quantum efficiency of the detectors (~55%) and various losses in fibres, optical elements and EOMs (~5%)—is roughly 10% per detector, which is a standard figure in many multi-photon down-conversion experiments.

Received 6 July; accepted 11 October 2006.

1. Bennett, C. & DiVicenzo, D. Quantum information and computation. *Nature* **404,** 247–255 (2000).
2. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409,** 46–52 (2001).
3. Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402,** 390–393 (1999).
4. Briegel, H. J. & Raussendorf, R. Persistent entanglement in arrays of interacting particles. *Phys. Rev. Lett.* **86,** 910–913 (2001).
5. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
6. Raussendorf, R., Brown, D. E. & Briegel, H. J. The one-way quantum computer – a non-network model of quantum computation. *J. Mod. Opt.* **49,** 1299–1306 (2002).
7. Nielsen, M. Optical quantum computation using cluster states. *Phys. Rev. Lett.* **93,** 040503 (2004).
8. Aliferis, P. & Leung, D. Computation by measurements: A unifying picture. *Phys. Rev. A* **70,** 062314 (2004).
9. Walther, P. *et al.* Experimental one-way quantum computing. *Nature* **434,** 169–176 (2005).
10. Kiesel, N. *et al.* Experimental analysis of a four-qubit photon cluster state. *Phys. Rev. Lett.* **95,** 210502 (2005).
11. Zhang, A. N. *et al.* Experimental construction of optical multiqubit cluster states from Bell states. *Phys. Rev. A* **73,** 022330 (2006).
12. Pittman, T. B., Jacobs, B. C. & Franson, J. D. Demonstration of feed-forward control for linear optics quantum computation. *Phys. Rev. A* **66,** 052305 (2002).
13. Giacomini, S., Sciarrino, F., Lombardi, E. & DeMartini, F. Active teleportation of a quantum bit. *Phys. Rev. A* **66,** 030302 (2002).
14. Ursin, R. *et al.* Quantum teleportation link across the Danube. *Nature* **430,** 849 (2004).
15. Grover, L. K. Quantum mechanics helps in search for a needle in a haystack. *Phys. Rev. Lett.* **79,** 325–328 (1997).
16. White, A. G., James, D. F. V., Eberhard, P. H. & Kwiat, P. G. Nonmaximally entangled states: production, characterization, and utilization. *Phys. Rev. Lett.* **83,** 3103–3107 (1999).
17. James, D., Kwiat, P., Munro, W. & White, A. Measurement of qubits. *Phys. Rev. A* **64,** 052312 (2001).
18. Toth, G. & Gühne, O. Detecting genuine multipartite entanglement with two local measurements. *Phys. Rev. Lett.* **94,** 060501 (2005).
19. Riebe, M. *et al.* Deterministic quantum teleportation with atoms. *Nature* **429,** 734–737 (2004).
20. Barrett, M. D. *et al.* Deterministic quantum teleportation of atomic qubits. *Nature* **429,** 737–739 (2004).
21. Vandersypen, L. M. K. *et al.* Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature* **414,** 883–887 (2001).
22. Nielsen, M. & Chuang, I. L. *Quantum Compuation and Quantum Information* (Cambridge Univ. Press, Cambridge, UK, 2000).
23. Nielsen, M. Journal club notes on the cluster-state model of quantum computation. ⟨http://www.qinfo.org/qc-by-measurement/cluster-state.pdf⟩ (2003).
24. Coffman, V., Kundu, J. & Wootters, W. K. Distributed entanglement. *Phys. Rev. A* **61,** 052306 (2000).
25. Horodecki, R., Horodecki, P. & Horodecki, M. Violating Bell inequality by mixed spin-1/2 states: necessary and sufficient condition. *Phys. Lett. A* **200,** 340–344 (1995).
26. Shor, P. W. in *Proc. 35th Annual Symp. on Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society Press, Los Alamitos, 1994).
27. Ahn, J., Weinacht, T. C. & Bucksbaum, P. H. Information storage and retrieval through quantum phase. *Science* **287,** 463–465 (2000).
28. Raussendorf, R., Browne, D. E. & Briegel, H. J. Measurement-based quantum computation on cluster states. *Phys. Rev. A* **68,** 022312 (2003).
29. Hein, M., Eisert, J. & Briegel, H. J. Multi-party entanglement in graph states. *Phys. Rev. A* **69,** 062311 (2004).
30. Soudagar, Y. *et al.* Cluster state quantum computing in optical fibres. Preprint at ⟨http://arxiv.org/quant-ph/0605111⟩ (2006).

Acknowledgements We are grateful to M. Aspelmeyer, Č. Brukner, J. I. Cirac, J. Kofler and K. Resch for discussions as well as to T. Bergmann and G. Mondl for assistance with the electronics. R.P. thanks E.-M. Röttger for assistance in the laboratory. We acknowledge financial support from the Austrian Science Fund (FWF), the European Commission under the Integrated Project Qubit Applications (QAP) funded by the IST directorate and the DTO-funded US Army Research Office.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.P. (robert.prevedel@univie.ac.at) or A.Z. (zeilinger-office@quantum.at).

LETTERS

Dilatant shear bands in solidifying metals

C. M. Gourlay[1] & A. K. Dahle[1,2]

Compacted granular materials expand in response to shear[1], and can exhibit different behaviour from that of the solids, liquids and gases of which they are composed. Application of the physics of granular materials has increased the understanding of avalanches[2], geological faults[3,4], flow in hoppers and silos[5], and soil mechanics[6,7]. During the equiaxed solidification of metallic alloys, there exists a range of solid fractions where the microstructure consists of a geometrically crowded disordered assembly of crystals saturated with liquid. It is therefore natural to ask if such a microstructure deforms as a granular material and what relevance this might have to solidification processing. Here we show that partially solidified alloys can exhibit the characteristics of a cohesionless granular material, including Reynolds' dilatancy[1] and strain localization in dilatant shear bands 7–18 mean crystals wide. We show that this behaviour is important in defect formation during high pressure die casting of Al and Mg alloys, a global industry that contributes over $7.3 billion to the USA's economy alone[8] and is used in the manufacture of products that include mobile-phone covers and steering wheels. More broadly, these findings highlight the potential to apply the principles and modelling approaches developed in granular mechanics to the field of solidification processing, and also indicate the possible benefits that might be gained from exploring and exploiting further synergies between these fields.

For most metallic alloys, the transformation from liquid to solid occurs gradually over a range of temperatures. This is accompanied by a continuous development in mechanical behaviour from a newtonian liquid to a polycrystalline viscoplastic solid, and an increase in apparent viscosity of 20 orders of magnitude[9]. Between these extremes, solid–liquid mixtures exist with rheological behaviour dependent on the deformation conditions and the partially solid microstructure. The microstructure is defined by the volume fraction, size, shape and distribution of the solid as well as by interfacial phenomena at crystal–crystal contacts. Interactions between crystals become important when solid comes into contact with solid owing to flow[10] or growth[11]—the crystallographic misorientation and solid–liquid interfacial energies determine whether a solid–solid interface or a liquid film is favourable. Additionally, for a given alloy, the microstructure can be varied by controlling the solidification or the deformation conditions and the microstructure is therefore strongly dependent on the thermomechanical history.

Past research on solid–liquid metallic mixtures can be broadly split into two areas. The first has concentrated on partially solid microstructures that have been engineered to tailor their rheological properties with the aim of controlling the flow behaviour during semisolid casting and forming processes[12]. Typically, this research has involved the formation of solid with globular morphology that exhibits reversible thixotropy under changing shear rate due to a dynamic equilibrium between neck growth and deformation at the welds between favourably oriented crystals[10,12].

In the second area, research has focused on the development of mechanical behaviour in the natural microstructures that evolve during equiaxed solidification. It has been shown that, after nucleation, crystals are initially dispersed in the liquid and the material behaves as a dilute suspension[13]. Crystal growth increases the volume fraction of solid (f_s) and, at a critical value of f_s, the crystals impinge on one another, causing a sharp increase in the resistance to flow due to the formation of a loose network of solid[14,15]. This f_s is often termed the dendrite coherency solid fraction (f_s^{Coh}) and is a strong function of the size and shape of crystals, ranging from ~0.15 for highly branched equiaxed dendrites, to ~0.5 for globular crystals[15]. At solid fractions slightly higher than f_s^{Coh}, the solid is able to transmit shear and compressive strains[13,14] but cannot transmit tensile strain[16]. Solidification beyond f_s^{Coh} increases the crowding and entanglement between crystals, and the shear strength increases approximately exponentially with solid fraction at constant shear rate[13]. As the solid has a homologous temperature of $T/T_s \approx 1$ (where T_s is the solidus temperature), both the solid and the liquid can be deformed and accommodate strain[17,18]. With continued solidification, the material develops partial cohesion between the solid[19], first owing to increased interlocking between irregularly shaped grains and, at higher solid fractions, owing to coalescence between grains through the formation of stable solid–solid interfaces[11]. Partial cohesion allows the solid to transmit tensile strain, and such microstructures have been mathematically described as partially cohesive viscoplastic porous media saturated with liquid[17,19]. These high solid fraction models[17] have successfully adopted ideas from soil mechanics, including Terzaghi's effective stress principle[20]. The rheology of the final stages of solidification has been extensively studied, as it is widely believed that these stages are critical in the formation of casting defects, including hot tears[21].

In the current study, we focus on shear behaviour at significantly lower solid fraction, in the range $f_s^{Coh} \leq f_s \leq 0.5$. Two hypoeutectic alloys have been used, Al-7Si-0.3Mg and Mg-9Al-0.7Zn, both of which solidify in an equiaxed mode over a wide range of solidification conditions and, in the range $0 < f_s \leq 0.5$, both alloys consist of a mixture of liquid and primary solid, L+(Al) or L+(Mg). In the first series of experiments, the Mg alloy was deformed during natural solidification using a four-bladed vane in a commercial rheometer. The sequence of events in a typical vane experiment is shown in Fig. 1a; expanded methods are given in Supplementary Information Section 1.2. It was found that experiments in which deformation was initiated at $f_s \geq f_s^{Coh}$ share the following characteristics: first, torque versus vane-rotation (M–θ) responses consist of an increase in torque to a peak value, followed by marked strain softening before deformation continues at lower torque (Fig. 1b). Second, we find that deformation is accompanied by a volumetric expansion: a maximum change in volume with rotation occurs at approximately the peak shear stress, and dilation then occurs more gradually with rotation during strain softening until an approximately constant volume is

[1]The CAST CRC, Materials Engineering, The University of Queensland, Brisbane, Queensland 4072, Australia. [2]The Australian Research Council CoE for design in light metals, Materials Engineering, The University of Queensland, Brisbane, Queensland, 4072, Australia.

reached (Fig. 1c). Third, examination of post-deformation samples reveals a band of concentrated porosity at the path circumscribed by the vane (Fig. 1d–f, and Supplementary Fig. 14), suggesting that the strain softening is associated with strain localization. In this Mg alloy, the shear band contains concentrated porosity at all cooling rates studied, and at cooling rates that result in large ($\gtrsim$700 μm) highly branched dendrites, crystal fragmentation was additionally observed in the band.

The shape of the M–θ and sample height versus vane-rotation (h–θ) responses in Fig. 1b, c and the localization of deformation into shear bands in Fig. 1d–f are typical characteristics of compacted cohesionless granular materials, such as dense sand or glass beads[7,22–26]. Compacted granular materials expand when sheared because particles must push one another apart and increase the space between themselves in order to rearrange (Reynolds dilatancy)[1,5]. The fact that partially solid alloys with $f_s > f_s^{Coh}$ exhibit similar behaviour indicates that, after growth has caused impingement ($f_s > f_s^{Coh}$), the crystals are sufficiently crowded that they cannot initially move past each other and that there is negligible intercrystal cohesion. The M–θ and h–θ responses suggest that crystals push one another apart in response to vane rotation, and that the dominant deformation mechanism is Reynolds dilatancy-enabled crystal rearrangement. The strain localization in Fig. 1d–f and Supplementary Fig. 14 can be explained by instabilities inherently caused by Reynolds dilatancy and fragmentation, because both decrease the local strength of the region in which they occur, promoting further deformation in that region.

Shear bands also form in direct shear cell experiments. An example is given for Al-7Si-0.3Mg with globular morphology in Fig. 2 and Supplementary Fig. 16, where a band exists on the shear plane containing a higher volume fraction of eutectic than adjacent regions (positive macrosegregation). As the eutectic was liquid at $f_s = 0.5$, the shear band had a higher liquid fraction than adjacent regions at the end of deformation, suggesting that liquid was drawn to the band during deformation.

We find that localized bands of porosity and positive macrosegregation form only when the material is deformed within a specific range of f_s: above those at which the material flows as a dilute suspension (when $f_s > f_s^{Coh}$) and below those at which the macroscopic shear response is crack propagation. This range of f_s is dependent on

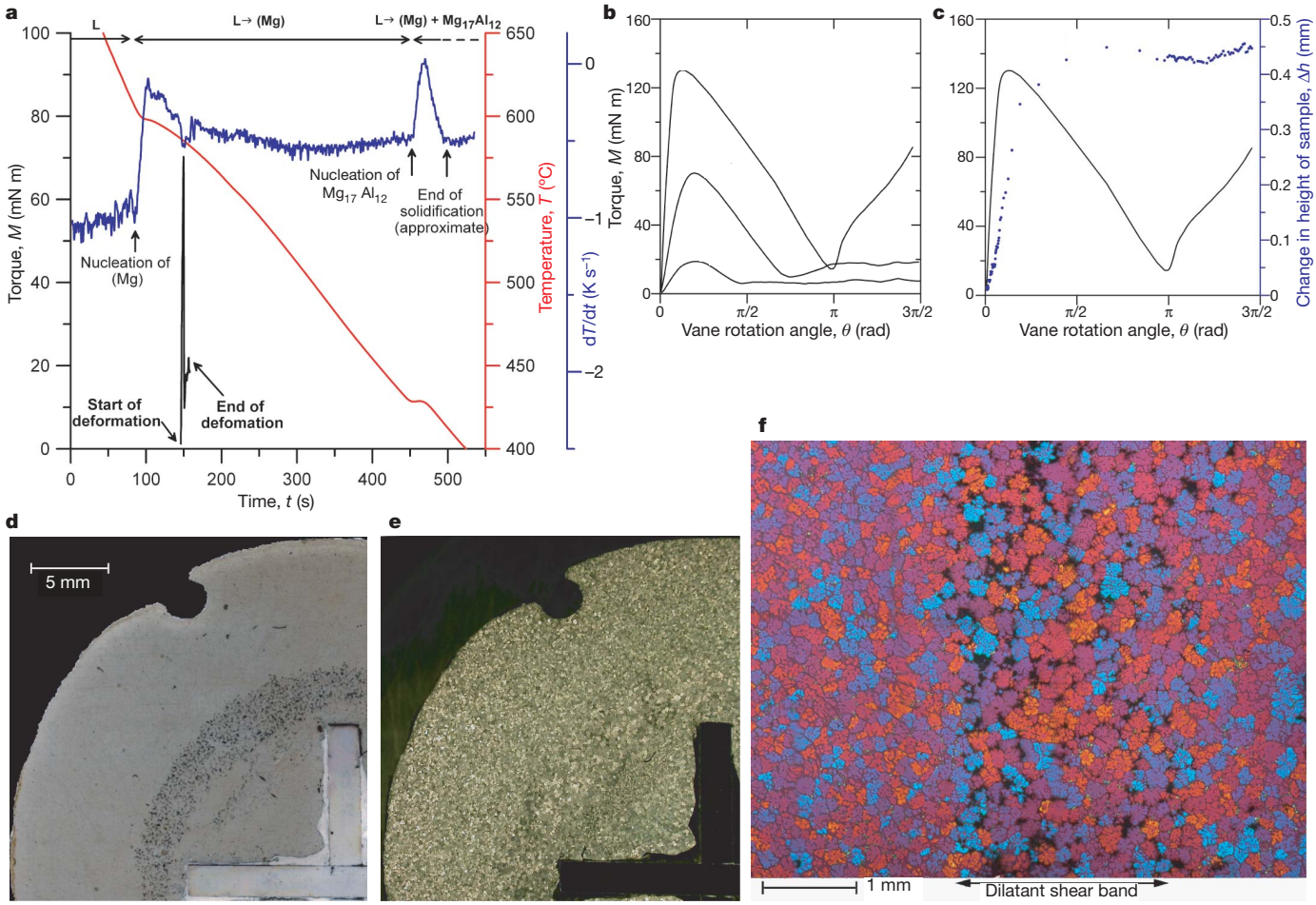


Figure 1 | Vane rheometry of partially solidified Mg-9Al-0.7Zn. a, Initially, liquid Mg-9Al-0.7Zn is cooled from 700 °C. Once nucleated, the decreasing temperature leads to dendritic growth of equiaxed (Mg) crystals. At a temperature corresponding to a desired f_s, the rotation of a four-bladed vane is initiated at 5 rotations per minute. Deformation continues for one vane rotation (12 s) and is then stopped. As the temperature continues to decrease, solidification progresses by the growth and coarsening of (Mg). At $f_s \approx 0.84$, the eutectic reaction L→(Mg) + $Mg_{17}Al_{12}$ commences. **b**, Torque–vane-rotation responses for three typical experiments conducted at $f_s > f_s^{Coh}$. Experimental parameters (temperature T in °C, f_s, and τ_{peak} in kPa) as follows: top curve (580.7, 0.35, 9.1); middle curve (585.5, 0.29, 4.9); bottom curve (590.2, 0.22, 1.3). **c**, A sample expands as it is sheared, and reaches a near-constant volume towards the end of deformation after global volumetric strain of $\varepsilon_{Vol} \approx 0.01$ ($T = 580.7$ °C, $f_s = 0.35$). **d**–**f**, Post-deformation microstructures of the Mg alloy after complete solidification in a sample deformed at $f_s = 0.19$. **d**, Macrograph of one-quarter of the cross-section through the centre of the vane. A localized band of porosity exists at the path circumscribed by the vane. **e**, The equiaxed grain structure throughout the same cross-section. **f**, A higher magnification image of the band shown in **d** and **e**, revealing that the porosity band is ~11 grains wide. The shear strain rate within shear bands is $\dot{\gamma} = 1$–$4\ s^{-1}$. A discussion of the changes in grain size and shape between deformation and complete solidification is given in Supplementary Information sections 1.7 and 2.1.3.

microstructural parameters, including the size and shape of crystals[18]. For deformation conditions similar to those used in this study, this range is typically $0.2 < f_s < 0.35$ for Al-7Si-0.3Mg with large branched dendrites[18]. The range is displaced to higher solid fraction for smaller, more compact dendrites[18], and globular microstructures exhibit a macroplastic response (without cracking) at $f_s > 0.5$ (Fig. 2). The effect of mechanical variables such as strain rate on this range is yet to be explored.

A wide variety of experimental studies have shown that dilatant shear bands in granular materials are 6–20 mean particles wide[7,22–28], and this has become known as a signature of granular deformation. To confirm whether the bands of porosity and macrosegregation in Figs 1 and 2 are dilatant shear bands, the relationship between shear band thickness (w) and grain size in the shear band (d_{sb}) was investigated. The grain size in the band was not always similar to the mean grain size in the sample because crystal fragmentation significantly decreases d_{sb}. In the vane and direct shear cell experiments, we find $w/d_{sb} = 7$–18 (Figs 1f, 3a, and Supplementary Figs 14–20), consistent with dilatant shear bands in granular materials (Fig. 3b). It therefore appears that the rheology of equiaxed partially solidified alloys with f_s slightly above f_s^{Coh} is similar to a cohesionless granular material such as dense sand. From a rheological perspective, the material is then a compacted, disordered assembly of metallic single crystals of homologous temperature ~1 with little intercrystal cohesion, saturated by a newtonian liquid of dynamic viscosity similar to distilled water (1.3×10^{-3} Pa s for Al at its melting point). This is an important result, as it suggests that the principles and modelling approaches developed through extensive experimental[1,7,22–26] and theoretical[1,4,6,22,27] studies on the mechanics of granular materials may be applicable to the less-studied field of partially solid alloy rheology and therefore to industrial casting processes.

One process where granular behaviour appears to be important is high pressure die casting (HPDC), during which large pressures (≤100 MPa) and stresses act on a solidifying alloy that consists of a large number of small crystals (~10 μm). Bands of defects form in HPDC components (Fig. 4 and Supplementary Fig. 19), which share common features with the rheometry experiments (Figs 1 and 2): in both rheometry and HPDC samples, the bands in the Mg alloy contain concentrated porosity (Figs 1d and 4c), the bands in the Al alloy contain an increased eutectic fraction (Figs 2b and 4a), and the macrosegregation profile in the HPDC Al alloy cross-sections follows the same trend as that after deformation in a direct shear cell (Figs 2c and 4b). In HPDC, the bands are 10–18 mean grains wide in both alloys, suggesting that defect bands in HPDC are also dilatant shear bands. Experimental studies on granular materials have found that w/d_{sb} is not a constant[7,23,24,26] but often varies with position in a sample[25] and is also a function of variables, including the confining pressure[26]. Despite this, over a broad range of conditions, w/d_{sb} is reported to be in the range 6–20 (refs 7, 22–27). Similarly, the deformation conditions in HPDC (which are largely unknown) are unlikely to be the same as in the rheometry experiments, yet in all experimental methods reported here, and for both alloys, w/d_{sb} lies within the range 7–18 (Fig. 3a). This discovery indicates that a granular approach to the modelling of filling and feeding in HPDC should be pursued.

We can apply the principles developed for granular materials to explain the formation of bands of macrosegregation and porosity in solidifying alloys: crystals are initially crowded, so that as the vane begins to rotate, the torque increases as the crystals push one another

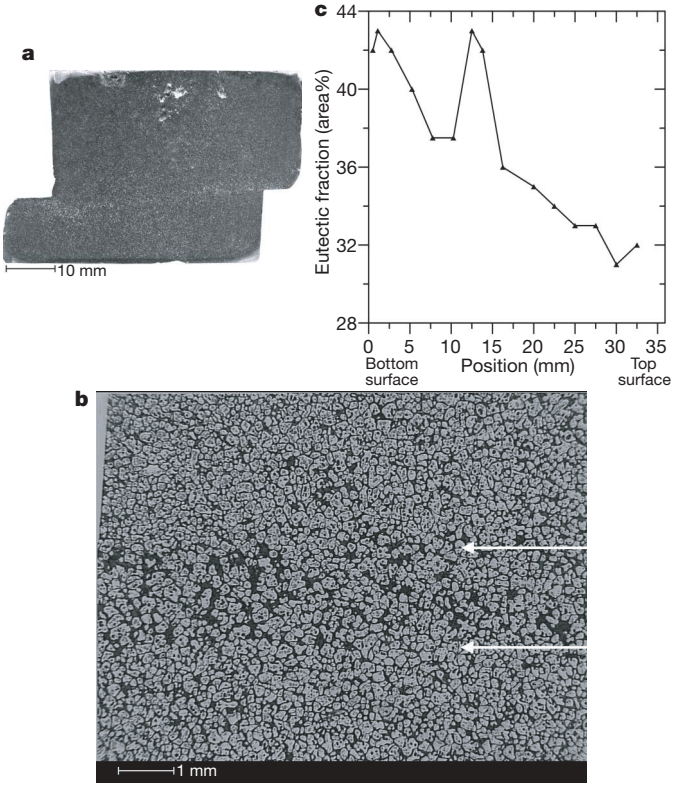


Figure 2 | Al-7Si-0.3Mg with globular morphology deformed in a direct shear cell. a, Cross-section through a sample deformed isothermally at $f_s = 0.5$ ($T = 580\ ^\circ C$) at 1.73 mm s^{-1} for 6 s. The sample was quenched at the end of deformation. See Supplementary Information section 1.3 for more detailed methods. **b**, An etched micrograph from the shear plane. Black material is [(Al)+Si] eutectic, which was liquid during deformation; grey material is (Al), which was mostly solid during deformation. A localized band of higher eutectic fraction (positive macrosegregation) exists on the shear plane, as indicated by white arrows. Average $\dot{\gamma} = 1\ s^{-1}$ within this shear band. **c**, Change in eutectic fraction with vertical position, as determined by image analysis. The sample exhibits a peak in eutectic fraction (and therefore liquid fraction during deformation) at the band shown in **b**. The general increase in eutectic fraction towards the bottom of the specimen is typical inverse segregation in the dominant direction of heat flow.

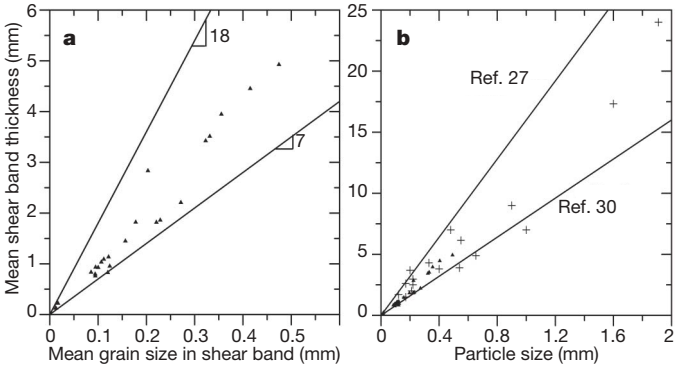


Figure 3 | Shear band thickness (w) and particle size in the shear band (d_{sb}). a, w/d_{sb} data from alloys. Shown are data from bands of porosity and macrosegregation created in the commercial casting process HPDC, and from three laboratory experiments: vane rheometry, a direct shear cell, and a gravity flow-through die casting technique. In all experiments, the bands are 7–18 mean grains thick. This w/d_{sb} relationship is observed over a wide range of grain sizes, facilitated by the markedly different cooling rates in the four methods, which ranged from ~1 to 1,000 K s^{-1}. This resulted in final grain sizes in the range ~1,000–10 μm. See Supplementary Information sections 1, 2.2 and 2.3 for further details of all experimental techniques and additional micrographs containing shear bands. **b**, Data from **a** re-plotted (triangles) with experimental results from the literature on the thickness of dilatant shear bands in common granular materials, such as dense sand and glass beads (crosses)[23–29]. Details of other researchers' data are given in Supplementary Table 5. The results of two theoretical studies that adopted different approaches to granular deformation are also shown[27,30]. The w/d_{sb} relationship for shear bands formed during equiaxed alloy solidification is consistent with that in granular materials.

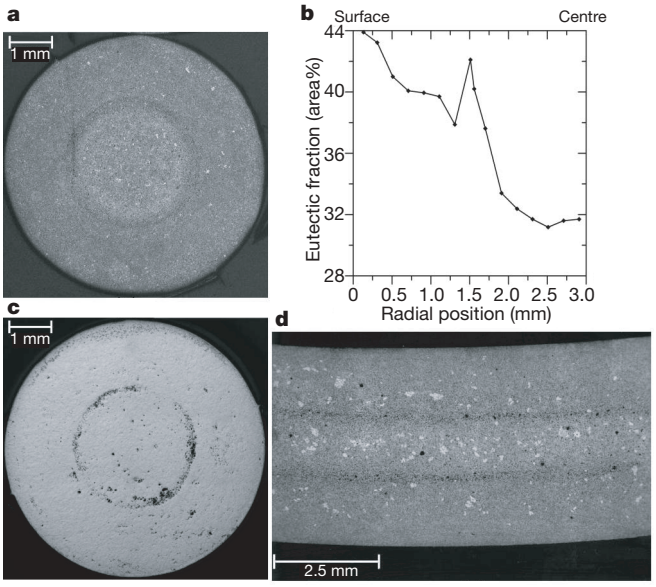


Figure 4 | Defect bands in high pressure die castings. a, c, Cross-sections through HPDC tensile bars in which the bulk flow direction during processing was into the page. **a,** HPDC Al-7Si-0.3Mg contains a dark band with $w = 236\ \mu m$ and $d_{sb} = 17\ \mu m$; **c,** HPDC Mg-9Al-0.7Zn exhibits a band of concentrated porosity with $w = 139\ \mu m$ and $d_{sb} = 13\ \mu m$. **b,** A macrosegregation profile from **a,** determined by image analysis, showing that the dark band is a local region of increased eutectic fraction (positive macrosegregation). **d,** Section of a HPDC Mg-5Al-0.2Mn commercial steering wheel, exhibiting bands containing pores and macrosegregation. The bulk flow direction during processing was from left to right.

apart and the material expands. As the crystal assembly dilates, strain instabilities, inherently caused by the changing volume, start to localize the deformation. During strain softening, both shear and dilatancy become concentrated in a shear band, which continues to dilate under shear towards a band microstructure that can flow at constant shear stress and without further expansion—a condition termed 'the critical state' in soil mechanics[24]. In saturated granular materials, liquid can be drawn to the expanding particle assembly, creating a band of increased liquid fraction[6]. After band formation, we suggest that the final appearance of a dilatant shear band is determined by post-deformation solidification (Fig. 1a). The formation of a band of higher liquid fraction affects the subsequent solidification in two ways: first, the increased liquid fraction increases the mean solute content of the band, because solute has been rejected into the liquid at solid–liquid interfaces of both alloys. Post-deformation solidification therefore results in positive macrosegregation. Second, the altered solid distribution can affect porosity formation: the liquid drawn to the band has a solute content similar to adjacent liquid at that time. Cooled at the same rate as adjacent regions after deformation, the liquid in the band would solidify at the same rate ($\partial f_s/\partial t$) as adjacent liquid. The higher liquid fraction in the band is therefore maintained as solidification progresses, and the band reaches the end of solidification later than adjacent regions. The band is therefore prone to contain porosity in alloys (such as Mg-9Al-0.7Zn) where solidification shrinkage is not adequately compensated for late in the solidification process.

Received 23 May; accepted 3 November 2006.

1. Reynolds, O. On the dilatancy of media composed of rigid particles in contact. *Phil. Mag.* **20,** 469–481 (1885).
2. Daerr, A. & Douady, S. Two types of avalanche behaviour in granular media. *Nature* **399,** 241–243 (1999).
3. Marone, C. & Kilgore, B. Scaling of the critical slip distance for seismic faulting with shear strain in fault zones. *Nature* **362,** 618–621 (1993).
4. Scott, D. Seismicity and stress rotation in a granular model of the brittle crust. *Nature* **381,** 592–595 (1996).
5. Duran, J. *Sands, Powders and Grains: An Introduction to the Physics of Granular Materials* (Springer, New York, 2000).
6. Vardoulakis, I. Deformation of water saturated sand: II. Effect of pore water flow and shear banding. *Geotechnique* **46,** 457–472 (1996).
7. Desrues, J. & Viggiani, G. Strain localization in sand: an overview of the experimental results obtained in Grenoble using stereophotogrammetry. *Int. J. Numer. Anal. Methods Geomech.* **28,** 279–321 (2004).
8. Shaping America's future. ⟨http://www.diecasting.org/information/dc_shape.htm⟩ (2005).
9. Kurz, W. & Fisher, D. J. *Fundamentals of Solidification* (Trans Tech Publications, Zurich, Switzerland, 1998).
10. Martin, C. L., Kumar, P. & Brown, S. Constitutive modeling and characterization of the flow behavior of semisolid metal alloy slurries. 2. Structural evolution under shear deformation. *Acta Metall. Mater.* **42,** 3603–3614 (1994).
11. Rappaz, M., Jacot, A. & Boettinger, W. J. Last-stage solidification of alloys: Theoretical model of dendrite-arm and grain coalescence. *Metall. Mater. Trans. A* **34,** 467–479 (2003).
12. Flemings, M. C. Behavior of metal alloys in the semisolid state. *Metall. Trans. A* **22,** 957–976 (1991).
13. Dahle, A. K. & Arnberg, L. Development of strength in solidifying aluminium alloys. *Acta Mater.* **45,** 547–559 (1997).
14. Metz, S. A. & Flemings, M. C. Hot tearing in cast metals. *AFS Trans.* **77,** 329–334 (1969).
15. Arnberg, L., Chai, G. & Bäckerud, L. Determination of dendritic coherency in solidifying melts by rheological measurements. *Mater. Sci. Eng. A* **173,** 101–103 (1993).
16. Stangeland, A., Mo, A., Nielsen, Ø., Eskin, D. G. & M'Hamdi, M. Development of thermal strain in the coherent mushy zone during solidification of aluminium alloys. *Metall. Mater. Trans. A* **35,** 2903–2915 (2004).
17. Martin, C. L., Favier, D. & Suéry, M. Viscoplastic behaviour of porous metallic materials saturated with liquid. Part I: constitutive equations. *Int. J. Plast.* **13,** 215–235 (1997).
18. Sumitomo, T., StJohn, D. H. & Steinberg, T. The shear behaviour of partially solidified Al-Si- Cu alloys. *Mater. Sci. Eng. A* **289,** 18–29 (2000).
19. Martin, C. L., Braccini, M. & Suéry, M. Rheological behavior of the mushy zone at small strains. *Mater. Sci. Eng. A* **325,** 292–301 (2002).
20. Terzaghi, K. *Theoretical Soil Mechanics* (Wiley & Sons, New York, 1943).
21. Eskin, D. G., Suyitno & Katgerman, L. Mechanical properties in the semi-solid state and hot tearing of aluminium alloys. *Prog. Mater. Sci.* **49,** 629–711 (2004).
22. Roscoe, K. H. The influence of strains in soil mechanics. *Geotechnique* **20,** 129–170 (1970).
23. Stephens, D. J. & Bridgwater, J. The mixing and segregation of cohesionless particulate materials. 1. Failure zone formation. *Powder Technol.* **21,** 17–28 (1978).
24. Muir Wood, D. Some observations of volumetric instabilities in soils. *Int. J. Solids Struct.* **39,** 3429–3449 (2002).
25. Oda, M. & Kazama, H. Microstructure of shear bands and its relation to the mechanisms of dilatancy and failure of dense granular soils. *Geotechnique* **48,** 465–481 (1998).
26. Wong, R. C. K. Shear deformation of locked sand in triaxial compression. *Geotech. Test. J.* **23,** 158–170 (2000).
27. Mühlhaus, H. B. & Vardoulakis, I. The thickness of shear bands in granular materials. *Geotechnique* **37,** 271–283 (1987).
28. Nemat-Nasser, S. & Okada, N. Radiographic and microscopic observation of shear bands in granular materials. *Geotechnique* **51,** 753–765 (2001).
29. Alshibli, K. A. & Sture, S. Sand shear band thickness measurements by digital imaging techniques. *J. Comput. Civ. Eng.* **13,** 103–109 (1999).
30. Bridgwater, J. On the width of failure zones. *Geotechnique* **30,** 533–536 (1980).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. J. Davidson for help recording the volumetric strain, H. Wang for providing a direct shear cell sample, Hydro Aluminium for providing the HPDC samples, and H. I. Laukli for discussions on defect bands in HPDC. This work was supported by the CAST CRC.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.M.G. (c.gourlay@minmet.uq.edu.au) or A.K.D. (a.dahle@minmet.uq.edu.au).

LETTERS

Influence of the intertropical convergence zone on the East Asian monsoon

Gergana Yancheva[1], Norbert R. Nowaczyk[1], Jens Mingram[1], Peter Dulski[1], Georg Schettler[1], Jörg F. W. Negendank[1], Jiaqi Liu[2], Daniel M. Sigman[3], Larry C. Peterson[4] & Gerald H. Haug[1]

The Asian–Australian monsoon is an important component of the Earth's climate system that influences the societal and economic activity of roughly half the world's population. The past strength of the rain-bearing East Asian summer monsoon can be reconstructed with archives such as cave deposits[1–3], but the winter monsoon has no such signature in the hydrological cycle and has thus proved difficult to reconstruct. Here we present high-resolution records of the magnetic properties and the titanium content of the sediments of Lake Huguang Maar in coastal southeast China over the past 16,000 years, which we use as proxies for the strength of the winter monsoon winds. We find evidence for stronger winter monsoon winds before the Bølling–Allerød warming, during the Younger Dryas episode and during the middle and late Holocene, when cave stalagmites suggest weaker summer monsoons[1–3]. We conclude that this anticorrelation is best explained by migrations in the intertropical convergence zone. Similar migrations of the intertropical convergence zone have been observed in Central America for the period AD 700 to 900 (refs 4–6), suggesting global climatic changes at that time. From the coincidence in timing, we suggest that these migrations in the tropical rain belt could have contributed to the declines of both the Tang dynasty in China and the Classic Maya in Central America.

Instrumental and historical records reaching back several centuries show considerable interannual to decadal variability in monsoonal strength. Although El Niño warm events tend to weaken the summer rainfall in both the Indian and East Asian monsoons[7], the instrumental data do not reveal a straightforward relationship among these major climate elements[8]. Palaeoclimate records can provide additional

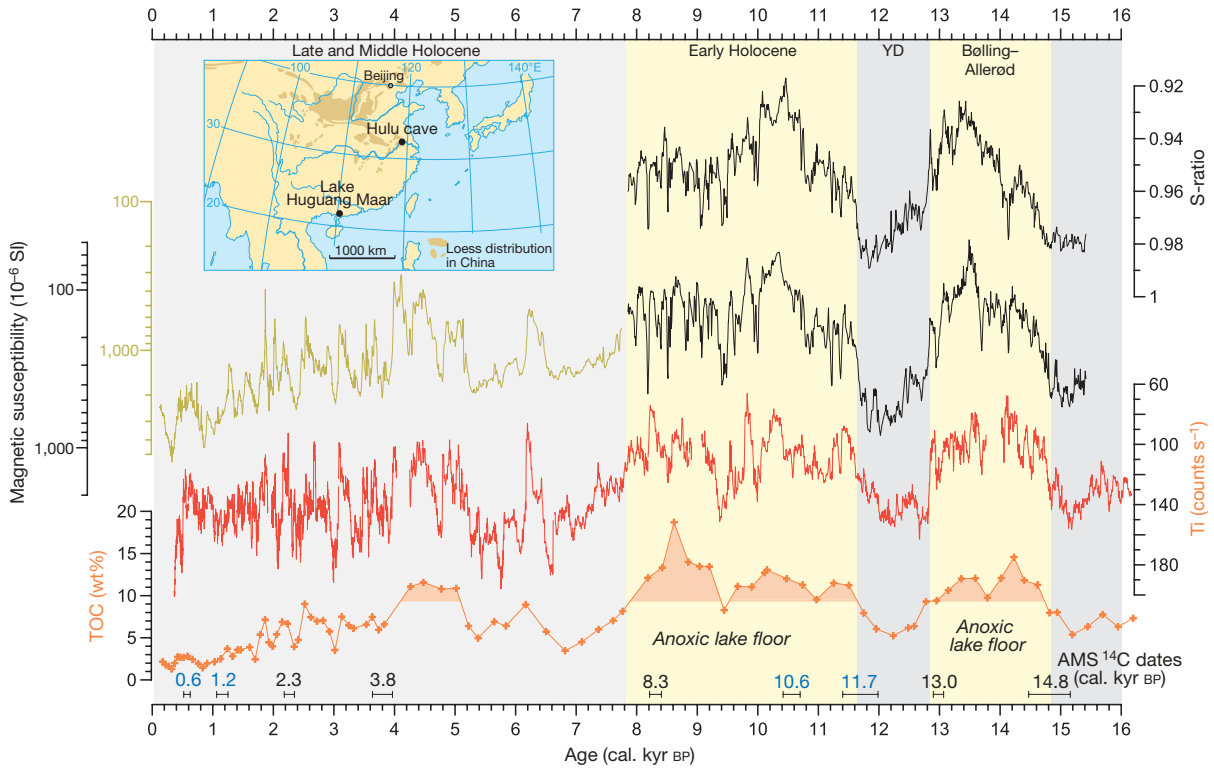


Figure 1 | Palaeoclimate time series of Lake Huguang Maar. Rock magnetic parameters (magnetic susceptibility and S-ratio), and Ti and TOC content from the sediment sequence during the past 16 kyr. Distinct intervals of anoxic conditions at the lake bottom are shaded in yellow. The S-ratio is saturated at 1 after 7.8 kyr BP. Five AMS ^{14}C dates of leaves (black) and four of bulk sediment (blue) are shown with an uncertainty interval of 2σ. Inset, locations of Lake Huguang Maar, Hulu cave and the Chinese loess plateau.

[1]GeoForschungsZentrum (GFZ), Section 3.3, Telegrafenberg, Potsdam D-14473, Germany. [2]Institute of Geology and Geophysics, Chinese Academy of Sciences, PO Box 9825, Beijing 100029, China. [3]Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA. [4]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida 33149, USA.

insights into monsoon dynamics, as the past includes larger amplitudes of climate change that may reveal more robust linkages. Previous palaeoclimate reconstructions generally agree that the Asian summer monsoon was weaker during cold phases in the Northern Hemisphere[1–3,9–15], when the intertropical convergence zone (ITCZ) tends to move southward[16–18], as it does during El Niño years[19–21].

Here we present a new palaeoclimatic record with nearly annual time-resolution from a sediment core recovered in Lake Huguang Maar, southeast China (21° 9′ N, 110° 17′ E), which extends back to 16.2 kyr ago (Fig. 1). The sedimentation rates range from 41 cm kyr^{-1} before the Bølling–Allerød to 112 cm kyr^{-1} during the past 4,000 yr. The age model is based on 5 AMS (accelerator mass spectrometry) ^{14}C dates of leaves and 4 of bulk sediment, with dating errors of less than ±160 yr within the 1σ interval of the AMS ^{14}C method. Adjustments using the well-dated records from Cariaco basin are within the error of the original ^{14}C-based age model (Fig. 1; see also Supplementary Information). Lake Huguang Maar today lies 23 m above sea level and has a water depth of 20 m. The surface area of the lake is 2.25 km^2, and it drains an extremely small catchment of 3.2 km^2. Because of its small catchment and a lack of stream inputs, the lake receives a minimal quantity of material by runoff and thus acts as a natural sediment trap for dust delivered to the site by the northerly winds of the winter monsoon. The Huguang Maar sediments record the strength of the winter monsoon in two independent ways: (1) the accumulation of wind-blown material, and (2) the redox-sensitive characteristics and total organic carbon (TOC) content of the sediment as a result of changes in wind stress and water-column mixing.

Our palaeoclimate time series are based on continuous measurements of sediment elemental composition and magnetic susceptibility, augmented by discrete measurements of additional magnetic properties and TOC content (Fig. 1). Micro X-ray fluorescence element scanning was performed at a resolution of 0.5 mm, smoothed with a 9 point triangular window. Magnetic susceptibility was measured in 2.5 mm steps directly on the surface of the split core halves. Additionally, sediment slices of 4 mm thickness (for the time interval of Termination I and the early Holocene) were analysed for their rock magnetic properties (see Methods).

Two rock magnetic parameters, magnetic susceptibility and the S-ratio (Fig. 1), measure the concentration of magnetic minerals and the mean oxidation state of iron in those minerals, respectively. The S-ratio is a (nonlinear) estimate of the abundance of magnetite compared to that of antiferromagnetic minerals, mainly haematite[22]. In Lake Huguang Maar, a high S-ratio indicates the availability of bottom water oxygen, and it is interpreted to reflect wind-driven lake mixing. Magnetic susceptibility is sensitive to both lake redox conditions and the aeolian input, both of which are affected by wind strength (see Methods).

The Ti content of the sediment is used to reconstruct the aeolian input into the lake (Fig. 1). The main lithogenic source to Huguang Maar sediments is dust transported by the winter monsoon winds from the arid areas in the north—for example, the loess plateau—and possibly local sources. Changes in Ti are interpreted as a measure of winter monsoon winds, with stronger winds more effectively transporting dense Ti- and Fe-rich grains (including the magnetic minerals) over the lake. Ti (rather than Fe) is used here as the dust input indicator because of its lack of redox sensitivity; however, the two are highly correlated (data not shown).

During cold climates, for instance, before 14.8 kyr ago and during the Younger Dryas (between 12.8 and 11.6 kyr ago), Ti content, magnetic susceptibility and S-ratio are high (Figs 1, 2) while TOC content is low[23]. In contrast, during the Bølling–Allerød and early Holocene (before 7.8 kyr ago), Ti content and rock magnetic amplitudes drop, and TOC increases (Fig. 1). Over the course of the Holocene, magnetic susceptibility clearly increases while the Ti content shows a weaker trend towards higher values; over this same time interval, the TOC content decreases (Fig. 1).

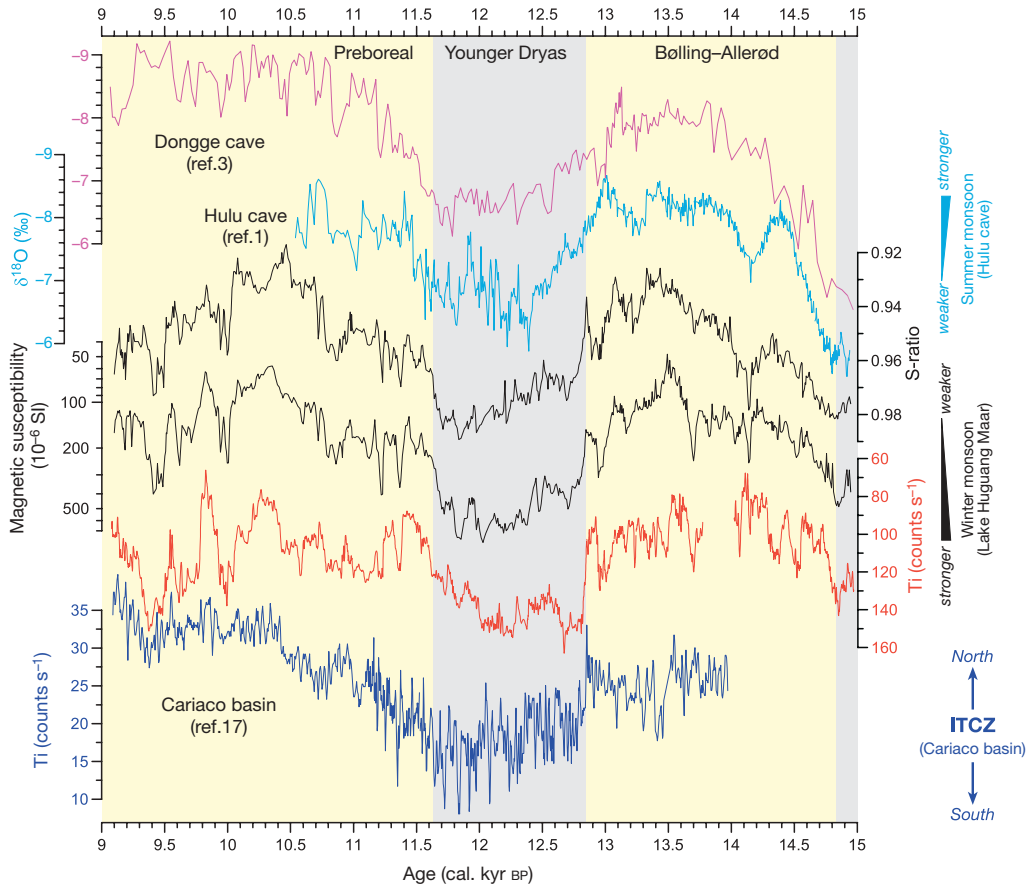


Figure 2 | Comparison of the monsoon sensitive sedimentary records from Lake Huguang Maar with other climate records. These are from the Cariaco basin[17], in the southern Caribbean off Venezuela, and Hulu and Dongge caves[1–3]. The Bølling-Allerød, Younger Dryas and Preboreal are highlighted.

The sediment S-ratio and TOC content, which indicate changes in bottom water oxygen, are functionally independent from Ti content, an indicator of dust input. Thus, their inter-consistency, reflecting high lake mixing whenever Ti-rich dust input is high, makes a compelling case for interpreting these records as a robust measure of winter monsoon strength. We deduce that the pre-Bølling–Allerød, the Younger Dryas, and the later Holocene were all characterized by strong winter monsoons in East Asia. During those intervals, low TOC content and higher magnetic susceptibility and S-ratios indicate enhanced wind mixing of the lake's waters, resulting in a well oxygenated lake floor, good preservation of magnetic minerals, and increased degradation of organic matter. In parallel, the higher Ti content during these times suggests intensified winter winds in central China leading to an increased input of dense Ti-bearing dust. In contrast, during periods of warmer climate such as the Bølling–Allerød and the early Holocene, Ti concentrations are lower, while lower magnetic susceptibility and S-ratios are coupled with higher TOC content, implying reduced winter monsoon winds and stable stratification of the lake.

The Lake Huguang Maar records of winter monsoon strength show a remarkable relationship with the radiometrically dated $\delta^{18}O$ records from stalagmites at Dongge cave[2,3] and Hulu cave[1], East Asia (Fig. 2). Strong winter monsoon phases reconstructed from Huguang Maar correlate with higher $\delta^{18}O$ at Hulu and Dongge caves, which indicates lower summer precipitation rates and thus a weaker summer monsoon[1]. Thus, our data, when compared with the speleothem[1,2] and South China Sea[24] records, argue for an inverse correlation between the strengths of the winter and summer monsoons (see also Supplementary Information). Moreover, the similarity of the records implies that monsoon changes during the latest glacial, Bølling–Allerød, Younger Dryas and Holocene were synchronous and common to large regions of coastal Southeast Asia. The inverse correlation between summer and winter monsoons recognized here is also evident for the Indian monsoon system, with the Younger Dryas representing a time of weaker summer and stronger winter monsoons[10]. The high-resolution records from East Asia indicate that this summer/winter monsoon anti-correlation also applies on shorter timescales (Fig. 2).

Migration of the annual mean position of the ITCZ provides a single coherent explanation for the observed trends in both winter and summer monsoons over the past 16 kyr, as well as for the strong anti-correlation between them. When the ITCZ is displaced northward, the summer monsoon should strengthen, while the winter monsoon will weaken. A northward shift in the ITCZ would be expected during times of Northern Hemisphere warming[25], such as the Bølling–Allerød and the early Holocene, times when, indeed, the East Asian summer monsoon was strong[1,26] and the winter monsoon was weak. This mode of explanation for climate change has recently proven fruitful for interpreting palaeoclimate records from the tropical Americas[17,18] (Fig. 2; see also Supplementary Information). The data reported here, in concert with existing data[1–3,20,21,27], suggest that these ITCZ migrations extended across the Pacific.

The role of climate and environmental change in the success or failure of societies is a matter of intense debate[4,5,28,29]. It would be simplistic to imagine that all episodes of societal change are driven by climatic events, especially in an advanced and complex society such as dynastic China. Nevertheless, we note that, on the basis of our new Huguang Maar data, major changes in Chinese dynasties[30] occurred when the winter monsoon was strong (Fig. 3). The anti-correlation between winter and summer monsoon strength indicated by comparison of the Huguang Maar data with the cave records would suggest that dynastic transitions tended to occur when the summer monsoon was weak and rainfall was reduced. Dynastic changes in China often involved popular uprisings during phases of crop failure

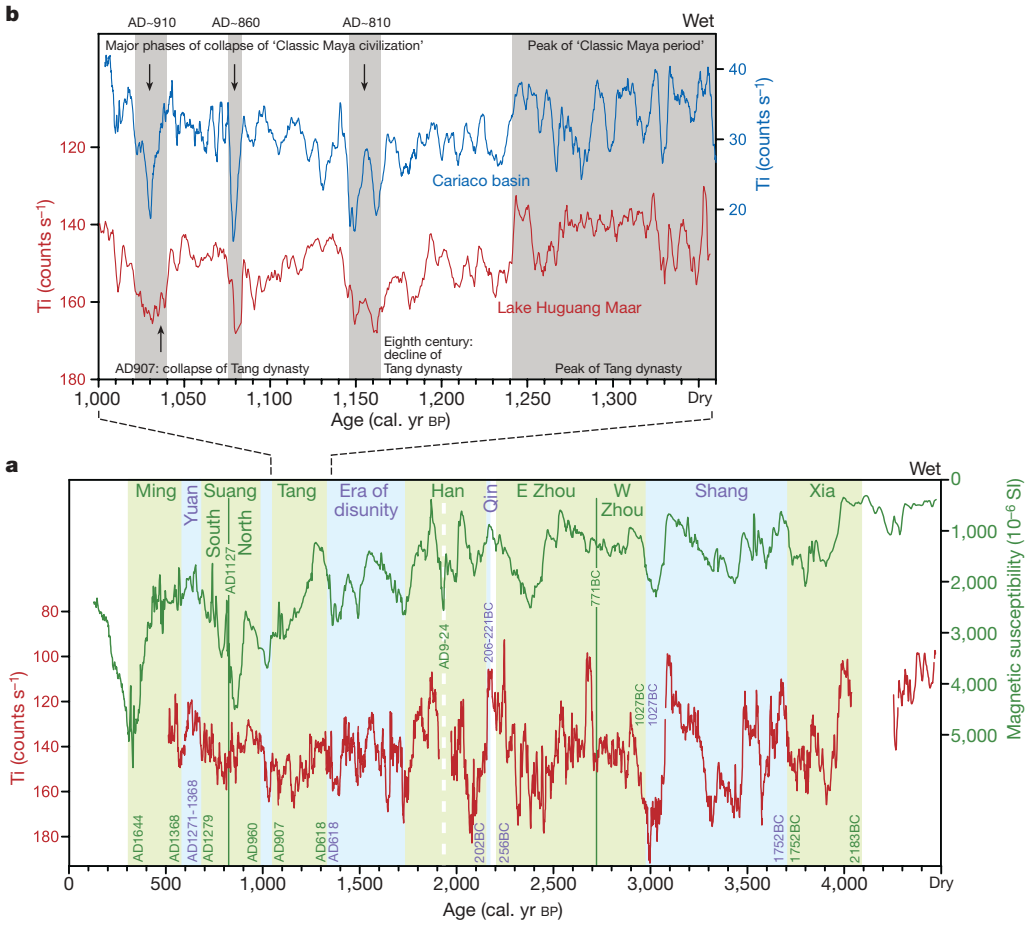


Figure 3 | The Lake Huguang Maar palaeoclimate records during the past 4,500 yr in the context of major events in the cultural history of China. **a**, Major changes in Chinese dynasties[30] occurred during dry phases, as indicated by the titanium and magnetic susceptibility records from Lake Huguang Maar and applying the observed anti-correlation between the winter and summer monsoons, while the described 'golden ages'[30] tended to occur during wet phases. **b**, Comparison of titanium records from Lake Huguang Maar and the Cariaco basin. The shared features of the two climate records as well as the similar timing of Chinese[30] and Mayan[4,5] societal changes suggest a role for coherent climate changes (that is, ITCZ migration) across the Pacific in the events of widely dispersed civilizations.

and famine, consistent with a linkage to reduced rainfall. The Tang dynasty has been described as a high point in Chinese civilization[30], a golden age of literature and art. The power of the dynasty began to ebb in the eighth century, starting with a defeat by the Arab army in AD 751. Rebellions further weakened the Tang empire, and it fully collapsed in AD 907 (ref. 30).

It is intriguing that the rise and collapse of the Classic Maya[4,5] coincided with the golden age and decline of the Tang dynasty in China[30]. Comparison of the Ti records from Lake Huguang Maar and the Cariaco basin reveals similarities, including both a general shift towards drier climate at about AD 750 and a series of three multi-year rainfall minima within that generally dry period (Fig. 3), the last of which coincides with the final stage of Maya collapse as well as the end of the Tang dynasty. Given these results, it seems possible that major circum-Pacific shifts in ITCZ position catalysed simultaneous events in civilizations on opposite sides of the Pacific Ocean.

METHODS

X-ray and magnetic measurements. Element scanning was carried out with a micro X-ray fluorescence spectrometer EAGLE BKA (Röntgenanalytik Meßtechnik GmbH). Isothermal remanent magnetizations (IRM), saturation as well as back field, were imprinted with a 2G Enterprises pulse magnetizer in magnetic fields of 2 T and −0.3 T, respectively. All IRMs were measured on a Molyneux spinner magnetometer (Minispin). The S-ratios were calculated after ref. 22. Continuous logging of magnetic susceptibility was performed with a Bartington MS2E sensor directly on the surface of split core halves. Magnetic susceptibility of the discrete samples was measured with a Kappabridge KLY-3S (AGICO).

S-ratio and TOC. The S-ratio varies between 0 for pure haematite and 1 for pure magnetite[22]. A high S-ratio indicates a predominance of oxic sedimentary conditions and the preservation of magnetite[22]. The strong correlation between high S-ratio and low TOC in our record has two alternative interpretations: (1) the supply of oxygen to the sediments is high, or (2) the flux of organic matter to the lake bottom is low. In the pervasively eutrophic Lake Huguang Maar, large changes in productivity are unlikely and difficult to cause by climate change. Thus, the measured changes in S-ratio and TOC are most logically interpreted as the result of wind-driven lake mixing, with more mixing and thus higher S-ratio and lower TOC occurring during times of strong winter monsoon winds. This interpretation is supported by sedimentary manganese and biogenic opal concentrations and accumulation rates (see Supplementary Information).

Magnetic susceptibility. The sharp increase in magnetic susceptibility at ~7.8 kyr ago, which is simultaneous with a saturating increase in S-ratio and a decrease in TOC content, requires that increasing wind mixing caused a threshold in lake oxygen content to be crossed, such that magnetite is subsequently preserved in the sediments (Supplementary Fig. 4). However, the continued gradual increase in magnetic susceptibility over the mid- to late Holocene has two potential explanations. It may result from a continued increase in the annual mean oxygen content of the lake, due to a continued increase in wind mixing. In this case, the lack of change in S-ratio would be due to saturation of this index at a value close to 1 (Supplementary Fig. 4). Alternatively, the gradual increase in magnetic susceptibility may result from an increase in the aeolian delivery of magnetic minerals because of stronger winter winds, as magnetic minerals such as magnetite are much (roughly two times) denser than most aluminosilicates. The gradual decrease in TOC content through the mid- to late Holocene may indicate either an increasing wind-driven ventilation of the lake or progressively greater dilution by aeolian inputs. In either case—an increase in lake mixing and/or an increase in aeolian input—the combined data require a Holocene strengthening of the winter monsoon.

Received 27 January; accepted 6 November 2006.

1. Wang, Y. J. *et al.* A high-resolution absolute-dated Late Pleistocene monsoon record from Hulu Cave, China. *Science* **294,** 2345–2348 (2001).
2. Yuan, D. *et al.* Timing, duration, and transitions of the Last Interglacial Asian Monsoon. *Science* **304,** 575–578 (2004).
3. Dykoski, C. A. *et al.* A high-resolution, absolute-dated Holocene and deglacial Asian monsoon record from Dongge Cave, China. *Earth Planet. Sci. Lett.* **233,** 71–86 (2005).
4. Diamond, J. *Collapse* (Penguin, London, 2005).
5. Fagan, B. *Floods, Famines and Emperors: El Nino and the Fate of Civilizations* (Pimlico, London, 2000).
6. Haug, G. H. *et al.* Climate and the collapse of Maya civilization. *Science* **299,** 1731–1735 (2003).
7. Kumar, K. K., Rajagopalan, B. & Cane, M. A. On the weakening of the relationship between the Indian monsoon and ENSO. *Science* **284,** 2156–2159 (1999).
8. Wang, B. *The Asian Monsoon* (Springer, Berlin, 2006).
9. Thompson, L. G. *et al.* Tropical climate instability: the last glacial cycle from a Qinghai-Tibetan ice core. *Science* **276,** 1821–1825 (1997).
10. Sirocko, F., Garbe-Schonberg, D., McIntyre, A. & Molfino, B. Teleconnections between the subtropical monsoons and high-latitude climates during the last deglaciation. *Science* **272,** 526–529 (1996).
11. Heslop, D. *et al.* Sub-millennial scale variations in East Asian monsoon systems recorded by dust deposits from the North-Western Chinese loess plateau. *Phys. Chem. Earth* **24,** 785–792 (1999).
12. Porter, S. C. & An, Z. Correlation between climate events in the North Atlantic and China during the last glaciation. *Nature* **375,** 305–308 (1995).
13. Oppo, D. W. & Sun, Y. Amplitude and timing of sea-surface temperature change in the northern South China Sea: Dynamic link to the East Asian monsoon. *Geology* **33,** 785–788 (2005).
14. Liu, T. & Ding, Z. Chinese loess and the paleomonsoon. *Annu. Rev. Earth Planet. Sci.* **26,** 111–145 (1998).
15. Ding, Z., Rutter, N., Han, J. & Liu, T. A coupled environmental system formed at about 2.5 Ma in East Asia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **94,** 223–242 (1992).
16. Hughen, K. A., Overpeck, J. T., Peterson, L. C. & Trumbore, S. Rapid climate changes in the tropical Atlantic region during the last deglaciation. *Nature* **380,** 51–54 (1996).
17. Haug, G. H., Hughen, K. A., Sigman, D. M., Peterson, L. C. & Röhl, U. Southward migration of the Intertropical Convergence Zone through the Holocene. *Science* **293,** 1304–1308 (2001).
18. Wang, X. *et al.* Wet periods in northeastern Brazil over the past 210 kyr linked to distant climate anomalies. *Nature* **432,** 740–743 (2004).
19. Cane, M. A. The evolution of El Nino, past and future. *Earth Planet. Sci. Lett.* **230,** 227–240 (2005).
20. Koutavas, A. & Lynch-Stieglitz, J. Marchitto Jr, T. M. & Sachs, J. P. El Nino-like pattern in ice age tropical Pacific sea surface temperature. *Science* **297,** 226–230 (2002).
21. Ivanochko, T. S. *et al.* Variations in tropical convection as an amplifier of global climate change at the millennial scale. *Earth Planet. Sci. Lett.* **235,** 302–314 (2005).
22. Bloemendal, J., King, J. W., Hall, F. R. & Doh, S. J. Rock magnetism of Late Neogene and Pleistocene deep-sea sediments: Relationship to sediment source, diagenetic processes and sediment lithology. *J. Geophys. Res.* **97,** 4361–4375 (1992).
23. Mingram, J. *et al.* The Huguang maar lake – a high-resolution record of palaeoenvironmental and palaeoclimatic changes over the last 78,000 years from South China. *Quat. Int.* **122,** 85–107 (2004).
24. Wang, L. *et al.* East Asian monsoon climate during the Late Pleistocene: high-resolution sediment records from the South China Sea. *Mar. Geol.* **156,** 245–284 (1999).
25. Hastenrath, S. & Greischar, L. Circulation mechanisms related to northeast Brazil rainfall anomalies. *J. Geophys. Res.* **98,** 5093–5102 (1993).
26. Wang, L. *et al.* Holocene variations in Asian monsoon moisture: a bidecadal sediment record from South China Sea. *Geophys. Res. Lett.* **26,** 2889–2892 (1999).
27. Fleitmann, D. *et al.* Holocene forcing of the Indian Monsoon recorded in a stalagmite from Southern Oman. *Science* **300,** 1737–1739 (2003).
28. deMenocal, P. B. Cultural responses to climate change during the late Holocene. *Science* **292,** 667–673 (2001).
29. Hodell, D. A., Brenner, M., Curtis, J. H. & Guilderson, T. Solar forcing of drought frequency in the Maya lowlands. *Science* **292,** 1367–1370 (2001).
30. Blunden, C. & Elvin, M. *Cultural Atlas of China* (Checkmark Books, New York, 1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Cane, R. Seager, P. deMenocal and S. Clemens for discussions, comments and reviews. This work was supported by the Deutsche Forschungsgemeinschaft (DFG). D.M.S. and G.H.H. thank the Humboldt Foundation for support. D.M.S. was also supported by BP and Ford Motor Company through the Princeton Carbon Mitigation Initiative.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.H.H. (haug@gfz-potsdam.de).

LETTERS

High-level similarity of dentitions in carnivorans and rodents

Alistair R. Evans[1], Gregory P. Wilson[2], Mikael Fortelius[1,3] & Jukka Jernvall[1]

The study of mammalian evolution depends greatly on understanding the evolution of teeth and the relationship of tooth shape to diet. Links between gross tooth shape, function and diet have been proposed since antiquity, stretching from Aristotle[1] to Cuvier[2], Owen[3] and Osborn[4]. So far, however, the possibilities for exhaustive, quantitative comparisons between greatly different tooth shapes have been limited. Cat teeth and mouse teeth, for example, are fundamentally distinct in shape and structure as a result of independent evolutionary change over tens of millions of years[5]. There is difficulty in establishing homology between their tooth components or in summarizing their tooth shapes, yet both carnivorans and rodents possess a comparable spectrum of dietary specializations from animals to plants. Here we introduce homology-free techniques[6–8] to measure the phenotypic complexity of the three-dimensional shape of tooth crowns. In our geographic information systems (GIS) analysis of 441 teeth from 81 species of carnivorans and rodents, we show that the surface complexity of tooth crowns directly reflects the foods they consume. Moreover, the absolute values of dental complexity for individual dietary classes correspond between carnivorans and rodents, illustrating a high-level similarity between overall tooth shapes despite a lack of low-level similarity of specific tooth components. These results suggest that scale-independent forces have determined the high-level dental shape in lineages that are widely divergent in size, ecology and life history. This link between diet and phenotype will be useful for inferring the ecology of extinct species and illustrates the potential of fast-throughput, high-level analysis of the phenotype.

The overall difficulties in analysing phenotypes are in contrast with the increasing availability and efficiency of analysing genomes. Successful linking of the genotype to the phenotype requires powerful tools, or 'fast-throughput morphometrics', for screening phenotypes and identifying the relevant details of the phenotype under natural selection.

Our approach to this task develops three-dimensional shape analysis and builds on relating the amount of mechanical processing that the food requires to the gross tooth form. The direct functional demands on tooth design depend on the required degree of mechanical processing[9,10]. In turn, the degree of mechanical processing that is required depends, first, on the mass-specific metabolic requirements of the animal and, second, on the difficulty with which mechanical and chemical breakdown of different kinds of foods can be achieved. We predict that the processing capability of the tooth will increase over evolutionary time when either of these two factors increases. An effective way of increasing processing capability is to add features onto the teeth that allow more food to be divided in each occlusal stroke. If we view teeth as 'tools' for breaking down food[11], this is like adding extra tools to the tooth that function in food breakdown. This is similar in meaning to 'breakage sites'[10]. 'Dental complexity' is then any measure of the number of features, tools or breakage sites on a tooth.

The foods of mammals vary extensively in their requirements for mechanical processing. For instance, vertebrate muscle is relatively easily digested, and so does not need to be fractured into small pieces for digestion. In contrast, the microbial digestion employed by mammals that eat fibrous plants requires efficient and repeated dental processing[12]. Thus, the proportion of muscle and analogous tissues to that of fibrous plant material gives a rough indication of the demands of the cheek teeth for mechanical processing. Although our approach allows broad-scale comparisons, it is of course a considerable simplification of the real situation in teeth, where the effectiveness of each 'tool' will vary both with its own shape and with the precise physical properties of the foods.

To test our prediction we measured cheek tooth complexity in two mammalian groups, carnivorans and rodents. Despite substantial differences in body size, chewing direction and physiology, members of both mammalian groups have independently and repeatedly evolved different dietary specializations, covering most of the range from animal to plant foods. This breadth of dietary specialization in radiations that have been distinct for at least 65 million years[5] makes carnivorans and rodents both suitable and separate tests of the association between diet and dental morphology. Individual taxa were chosen for this study on the basis of the availability of detailed dietary information from the wild, phylogenetic position, and the availability of dental material. The sample included 32 species of carnivorans (Ailuridae, Canidae, Felidae, Herpestidae, Hyaenidae, Mustelidae, Procyonidae, Ursidae and Viverridae) and 49 species of murine, sigmodontine and otomyine rodents, the first two of these being commonly known as rats and mice of the Old and New Worlds, respectively. We concentrated mainly on murine and sigmodontine rodents because they represent a major component of recent rodent diversity, with the members showing disparate and independent specializations to different diets. All scans in this study are viewable in the MorphoBrowser database, a web-accessible database with an interactive three-dimensional viewer (see Methods). We used five dietary categories to classify the species in this study, in order of increasing processing demands on the teeth: hypercarnivore, carnivore (including insectivores), animal-dominated omnivore, plant-dominated omnivore, and herbivore (specifically stem and leaf feeders, composed of grazers, browsers and mixed feeders). Whereas both mammalian groups show considerable overlap in dietary specializations, rodents lack hypercarnivores (that is, dedicated vertebrate flesh eaters) and carnivorans have few taxa in the plant-dominated omnivore and herbivore categories.

We developed fast-throughput procedures for the three-dimensional analysis of teeth. Teeth were first digitized; this was followed by

[1]Evolution and Development Unit, Institute of Biotechnology, PO Box 56 (Viikinkaari 9) FIN-00014 University of Helsinki, Finland. [2]Department of Earth Sciences, Denver Museum of Nature and Science, 2001 Colorado Boulevard, Denver, Colorado 80205, USA. [3]Department of Geology, PO Box 64 (Gustaf Hällströmin katu 2a) FIN-00014 University of Helsinki, Finland.

processing of the data files to produce digital tooth shapes for the analyses. For efficient computation of the various GIS and other complexity measures, a custom computer program was written (available from the authors on request). Currently, the most time-intensive step is the three-dimensional scanning (1–3 h) and initial processing of data (less than 30 min) into the GIS format, whereas computations take only seconds for each tooth row. As scanning and data processing technologies are developing at a rapid rate, this approach will quickly become increasingly efficient.

Both upper and lower cheek tooth rows (carnassials and all molars present) were digitized with a high-resolution laser scanner, and the three-dimensional point files were converted into digital elevation models of the tooth rows (Fig. 1). We used whole cheek tooth rows because this gives a better estimation of the overall processing capacity than single teeth. The results can be normalized for the number of teeth in the row, but the basic pattern remains unchanged. Because we were interested in shape apart from size, all tooth rows were scaled to the same length. To approximate the number of 'tools' on the crown[11], we subdivided the surface of each digital elevation model into patches based first on slope orientation and then on topographic elevation. Orientation maps were generated by determining the orientation at each grid point on the topographic maps as being one of eight compass directions (for example north and southwest). The maps were divided into patches by grouping contiguous points on the same contour level or with the same orientation together as a 'patch'. Next we used three different methods for calculating dental complexity and information content for both the orientation and the topographic patches: orientation and topographic patch count (OPC and TPC), orientation and topographic patch diversity (a measure of Shannon information; OPD and TPD), and image compression ratio of surface maps (OIC and TIC; see Methods and Supplementary Information).

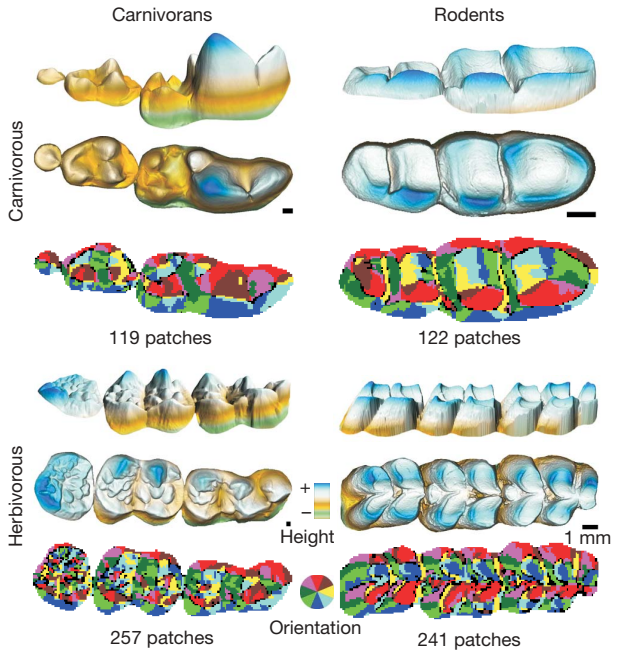


Figure 1 | Dental and dietary diversity in carnivorans and rodents. Three-dimensional buccal–occlusal and occlusal reconstructions of two carnivoran tooth rows (top left, red fox *Vulpes vulpes*; bottom left, giant panda *Ailuropoda melanoleuca*) and two rodent tooth rows (top right, golden-bellied water rat *Hydromys chrysogaster*; bottom right, Rothschild's woolly rat *Mallomys rothschildi*) for the GIS analysis. Determination of surface orientation (below each three-dimensional reconstruction, with orientation indicated by colour as shown on the colour wheel) allows the measurement of OPC (the number of coloured patches is indicated under each figure). These measures are compared with diets, namely carnivorous and herbivorous. Clumps smaller than three grid points are coloured black. Lower right tooth rows; anterior towards the right. Scale bars, 1 mm.

For carnivorans, OPC shows a relatively clear gradation in dental complexity from low values in hypercarnivores, intermediate in the omnivores, and highest in the herbivores (Fig. 2; $P < 0.001$ for all tests) in both the upper and lower tooth rows. Significant differences are also found between the dietary categories for TPC, OPD, TPD, OIC and TIC ($P < 0.05$ for all tests). The rodents illustrate a similar trend of dental complexity with diet (Fig. 2). Differences between the dietary categories were significant for OPC ($P < 0.01$ for lower tooth rows, $P < 0.05$ for upper tooth rows), but not for the other measures ($P > 0.05$ for all tests). We note that the better resolving power of OPC may be due to its identifying distinct functional surfaces (such as wear facets), fitting with the concept of tooth crown consisting of individual 'tools' for breaking down food[11].

When the same comparisons were made within the murines and the sigmodontines in the rodent sample, the OPC for the lower tooth row remained significantly different between the dietary categories ($P < 0.01$). This significant pattern of rodent OPC in relation to diet indicates that there remain selective pressures on molar form despite all rodents' having continuously growing incisors. Omnivore categories tend to have the largest ranges and are the least well resolved of the categories. This is perhaps expected, given that the diversity of foods that members of these two classes consume is likely to be much greater than at either end of the spectrum. Within the carnivorans, several of the families have a limited dietary range (for example, felids in hypercarnivores, and ursids in animal-dominated omnivores and herbivores), but for families that span several dietary categories (for example, mustelids) the patch count varies according to diet, indicating intrafamily and interfamily resolution.

Furthermore, when we compare the absolute values of the complexity measures between carnivorans and rodents, the ranges of dental complexity values in the two taxonomic groups overlap one another (Fig. 2). Only the carnivore diet category differs significantly between the two taxonomic groups (except for upper OPC ($P = 0.439$); Mann–Whitney *U*-tests; see Supplementary Information). This may reflect the fact that the carnivorous rodents are mostly invertebrate feeders, whereas the corresponding carnivorans are mostly vertebrate feeders. Indeed, if the carnivoran species that do not include moderate amounts of insects in their diet are excluded (*Crocuta crocuta*, *Gulo gulo*, *Mustela erminea*, *M. lutreola* and *M. putorius*), the mean OPC for the lower tooth row rises from 92 to 110, although this is still below that of carnivorous rodents (lower OPC mean 156, $P < 0.005$). The higher OPC of insectivores is likely to be linked to the greater development of shearing crests noted previously[13,14].

This high-level similarity between carnivoran and rodent dentitions reported here is noteworthy because the extensive differences in the low-level details of cheek tooth shape (number and position of cusps, crests and folds), number of teeth, tooth classes represented (muroids do not have canines or premolars), replacement (muroids do not have a deciduous dentition) and chewing motion[15,16] make it difficult to identify comparable landmarks and features in carnivoran and rodent teeth. In addition, body sizes, and consequently tooth sizes, are greatly different in the two groups, ranging from 5 g to 1 kg in rodents and from 90 g to 380 kg in carnivorans included in this sample. Despite these fundamental differences, our results show not only a similar tendency in relation to diet but also comparable complexity values in rodents and carnivorans for each dietary category. We interpret these results as being strongly indicative of scale-independent and phylogeny-independent effects of diet on general aspects of dental shape. It has been shown that tooth shapes can diverge relatively rapidly between populations as long as the occlusal fit is maintained[17,18], and it remains to be tested how evolutionarily labile OPC is with respect to other shape parameters.

Several anatomical and physiological characteristics have recently been shown to reflect the degree of carnivory and herbivory in disparate ranges of living mammals (for example, intracellular targeting of alanine:glyoxylate aminotransferase[19] and salivary gland

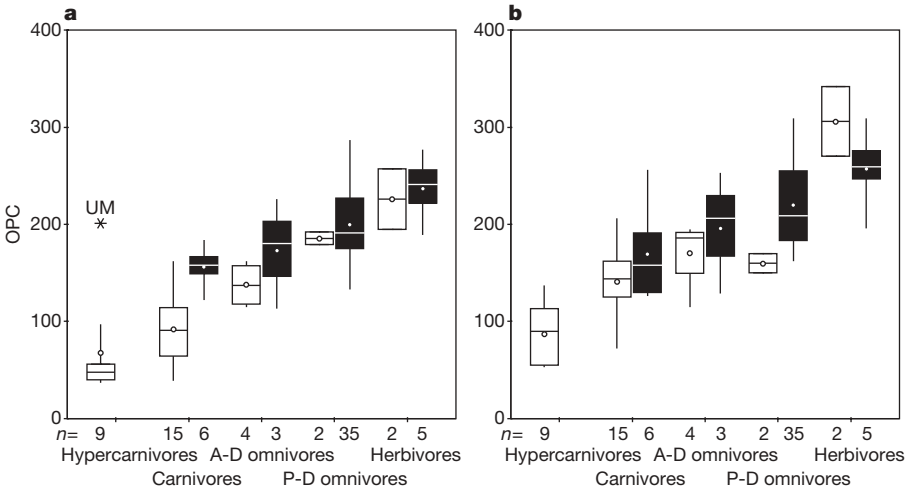


Figure 2 | Dental complexity follows diet similarly in carnivorans and rodents. Tooth complexity (measured as OPC) for five major dietary types in two taxonomically disparate groups (carnivorans (open boxes) and rodents (filled boxes)) for the lower (**a**) and upper (**b**) tooth rows. There is a consistent increase in the dental complexity when moving from hypercarnivory (meat-feeding) through animal-dominated (A-D) and plant-dominated (P-D) omnivory (meat and plants) to herbivory (plant material such as leaves and grasses). For several of the dietary categories, a similar range of values is found in both carnivorans and rodents (for example 'herbivores' in the lower tooth row and 'carnivores' in the upper tooth row). Boxes enclose 50% of observations; the median and mean are indicated with a horizontal bar and circle, respectively, and whiskers denote range, other than the one extreme outlier (UM, *Ursus maritimus*), which is indicated with an asterisk. *n*, number of species in each category.

structure[20]). However, it remains to be determined how quickly aspects of physiology and dental complexity respond to shift in diet. In our data the polar bear (*Ursus maritimus*) seems to be the main exception to the measured patterns. The relatively high dental complexity value of the highly carnivorous polar bear is likely to reflect its recent divergence from the brown bear (*Ursus arctos*), a plant-dominated omnivore, during the late Pleistocene (250–200 kyr ago), with morphological divergence perhaps only in the past 20–40 kyr (refs 21–23). Nevertheless, the upper tooth row of the polar bear still has lower dental complexity values than the other ursids in the sample (*U. arctos*, *U. americanus* and *Ailuropoda melanoleuca*), indicating that some change in the expected direction has already occurred at an extremely rapid rate. Furthermore, in comparison with other ursids, the polar bear has a substantially reduced relative tooth area (see Supplementary Information). We note, however, that whereas in our data there is a tendency towards larger tooth size in herbivores, dental complexity more completely differentiates the species according to diet and without the need for body size information (see Supplementary Information). We interpret these patterns of results to support the conclusion[10] that tooth-size–body-size predictions are best made on species groups with homologous diets.

Thus, we have shown here that in the evolution of two major mammalian groups, carnivorans and rodents, the consumption of more demanding foods has resulted in the evolution of more complex teeth in multiple independent lineages, which we could liken to the similarity of the sums irrespective of the parts. In combination with recent views of emergent properties of dental development[24], this strongly suggests that simple functional and developmental considerations may explain the bewildering diversity of tooth shapes observed in fossil and living mammals. Our results hold promise for the use of OPC in reconstructing diets of extinct taxa, even in cases where the use of living analogues is difficult because of disparate morphologies. In this respect our approach of using three-dimensional morphology is analogous to recent advances in determining three-dimensional texture of micro-wear in fossil hominins[25] and underscores the potential for fast-throughput data acquisition and analyses of living and fossil taxa.

METHODS

Diet categories. The modern species sampled from carnivorans and rodents were placed into five dietary categories (hypercarnivore, carnivore, animal-dominated omnivore, plant-dominated omnivore and herbivore) roughly reflecting the increasing demands of mechanical processing; however, for detailed discussion on mechanical properties of foods, see ref. 10. Diets were obtained from ref. 26 and monographic sources from the literature (see Supplementary Information). In this study, the 'herbivore' category is limited to stem and leaf feeders, which includes grazers, browsers and mixed feeders. Species with diets including substantial amounts of other plant material and occasional feeding on animals were placed in the 'plant-dominated omnivore' category. The wide variety of carnivorans (members of the order Carnivora) in the study cover the range from hypercarnivory (felids and some canids) through plant-dominated omnivory (bears) to herbivory (giant panda), amounting to 32 species. The rodents are largely represented by the murine and sigmodontine radiations, which account for more than 45% of rodent species diversity. Our sample includes 49 rodent species that cover a very wide dietary range.

Three-dimensional scans. One upper and one lower tooth row of each species were scanned with a Nextec Hawk three-dimensional laser scanner at between 10 and 50 μm resolution, depending on the size of the tooth row. Scans were entered into the MorphoBrowser database (http://morphobrowser.biocenter.helsinki.fi/). Teeth were oriented manually to maximize crown–base projection. For carnivorans, the carnassials (upper 4th premolar, P^4, and lower 1st molar, M_1) and all teeth posterior to them were scanned, representing between one tooth (a single M_1 in species such as *Felis silvestris*) and four teeth (P^4 to M^3 and M_1 to M_4 for *Otocyon megalotis*); for the rodents, the entire molar row, which is either two or three teeth in each jaw, was scanned. To standardize for size, each tooth row was represented by 150 data rows (typically less than half of the scanning resolution), varying in the number of columns depending on the relative width of the tooth row. To obtain functionally comparable measures of tooth shape, only carnivoran specimens with light wear were used, and for the rodents, which already have enamel-free areas on their unworn cusp tips, we standardized the wear state to a moderate level. Whereas specific functional features have been shown to be modified by tooth wear[7], general topographic measures are more stable[6,8], indicating that high-level patterns should be relatively robust to tooth wear.

Patch count, patch diversity and image compression. Initial interpolation of a regular grid of points was performed with Surfer for Windows (Golden Software, Inc.). Topographic (contour) maps were then generated with contours of twice the *x* and *y* resolutions. Using custom GIS software written by one of the authors (A.R.E.), the topographic and orientation maps were divided into patches, with a minimum patch size of three grid points. The number of these patches is the patch count for TPC and OPC maps, respectively (see Fig. 1 for example). To test for the effect of detected feature coarseness, variations in contour size (twice or four times *x* and *y* resolution, and *z* range divided by five or ten), orientation (four or eight orientations) and minimum patch size (3 or 11) were examined. To test for the effect of tooth orientation, surface with a slope of less than 5° or 10° was either voided or grouped as a separate patch. Except for some of the

coarsest levels, these variations did not affect the significance of the observed patterns (see Supplementary Information).

The following methods were used to estimate 'information content' or patch diversity of the tooth surface. TPD and OPD were calculated as $1/\sum[(\text{patch size in grid points})^2/(\text{total number of grid points})^2]$. This is a measure of information content derived from ref. 27, with the current method based on ref. 28. Each TIC and OIC map was compressed by using the JPEG and PNG algorithms, and the compression ratio was used as a measure of information content[29]. IrfanView (http://www.irfanview.com/) was used for image compression, using JPEG 80% quality (10% was also tested) and PNG compression level 6 (level 9 was also tested).

For all measures, statistical differences between dietary categories were tested with Kruskal–Wallis tests, and differences between taxonomic groups within dietary categories with the use of Mann–Whitney *U*-tests, each with a two-tailed Monte Carlo estimation of significance with 10,000 samples performed in SPSS version 11.0 (SPSS Inc.).

Received 4 September; accepted 9 November 2006.
Published online 13 December 2006.

1. Aristotle. *Parts of Animals* (transl. Peck, A. L.) (Harvard Univ. Press, Cambridge, Massachusetts, 1983).
2. Cuvier, G. *Discours sur les Révolutions de la Surface du Globe, et sur les Changements qu'elles ont Produits dans le Règne Animal* (Dufour et d'Ocagne, Paris, 1825).
3. Owen, R. *Odontography* (Hippolyte Baillière, London, 1840–1845.)
4. Osborn, H. F. *Evolution of Mammalian Teeth, to and from the Triangular Type* (Macmillan, New York, 1907).
5. Wible, J. R., Rougier, G. R. & Novacek, M. J. in *The Rise of Placental Mammals: Origins and Relationships of the Major Extant Clades* (eds Rose, K. D. & Archibald, J. D.) 15–36 (Johns Hopkins Univ. Press, Baltimore, Maryland, 2005).
6. Ungar, P. S. & M'Kirera, F. A solution to the worn tooth conundrum in primate functional anatomy. *Proc. Natl Acad. Sci. USA* **100,** 3874–3877 (2003).
7. Evans, A. R. Connecting morphology, function and tooth wear in microchiropterans. *Biol. J. Linn. Soc.* **85,** 81–96 (2005).
8. King, S. J. *et al.* Dental senescence in a long-lived primate links infant survival to rainfall. *Proc. Natl Acad. Sci. USA* **102,** 16579–16583 (2005).
9. Fortelius, M. in *Teeth Revisited: Proceedings of the VIIth International Symposium on Dental Morphology* (eds Russell, D. E., Santoro, J.-P. & Sigogneau-Russell, D.) 459–462 (Mémoires du Muséum national d'Histoire naturelle C, Paris, 1988).
10. Lucas, P. W. *Dental Functional Morphology* (Cambridge Univ. Press, Cambridge, 2004).
11. Evans, A. R. & Sanson, G. D. The tooth of perfection: functional and spatial constraints on mammalian tooth shape. *Biol. J. Linn. Soc.* **78,** 173–191 (2003).
12. Stevens, C. & Hume, I. *Comparative Physiology of the Vertebrate Digestive System* (Cambridge Univ. Press, Cambridge, 1995).
13. Kay, R. F. The functional adaptations of primate molar teeth. *Am. J. Phys. Anthropol.* **43,** 195–215 (1975).
14. Strait, S. G. Differences in occlusal morphology and molar size in frugivores and faunivores. *J. Hum. Evol.* **25,** 471–484 (1993).
15. Hillson, S. *Teeth* (Cambridge Univ. Press, Cambridge, 2005).
16. Peyer, B. *Comparative Odontology* (Univ. of Chicago Press, Chicago, Illinois, 1968).
17. Polly, P. D. On morphological clocks and paleophylogeography: towards a timescale for *Sorex* hybrid zones. *Genetica* **112,** 339–357 (2001).
18. Polly, P. D., Le Comber, S. C. & Burland, T. M. On the occlusal fit of tribosphenic molars: Are we underestimating species diversity in the Mesozoic? *J. Mamm. Evol.* **12,** 283–299 (2005).
19. Birdsey, G. M. *et al.* A comparative analysis of the evolutionary relationship between diet and enzyme targeting in bats, marsupials and other mammals. *Proc. R. Soc. B* **272,** 833–840 (2005).
20. Phillips, C. J., Weiss, A. & Tandler, B. Plasticity and patterns of evolution in mammalian salivary glands: comparative immunohistochemistry of lysozyme in bats. *Eur. J. Morphol.* **36,** 19–26 (1998).
21. Kurtén, B. The evolution of the polar bear, *Ursus maritimus* Phipps. *Acta Zool. Fenn.* **108,** 1–30 (1964).
22. Kurtén, B. *Pleistocene Mammals of Europe* (Weidenfeld & Nicolson, London, 1968).
23. Talbot, S. L. & Shields, G. F. Phylogeography of brown bears (*Ursus arctos*) of Alaska and paraphyly within the Ursidae. *Mol. Phylogenet. Evol.* **5,** 477–494 (1996).
24. Kangas, A. T., Evans, A. R., Thesleff, I. & Jernvall, J. Nonindependence of mammalian dental characters. *Nature* **432,** 211–214 (2004).
25. Scott, R. S. *et al.* Dental microwear texture analysis shows within-species diet variability in fossil hominins. *Nature* **436,** 693–695 (2005).
26. Nowak, R. *Walker's Mammals of the World* (Johns Hopkins Univ. Press, Baltimore, Maryland, 1999).
27. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 623–656 (1948).
28. Eterovick, P. C., Figueira, J. E. C. & Vasconcellos-Neto, J. Cryptic coloration and choice of escape microhabitats by grasshoppers (Orthoptera: Acrididae). *Biol. J. Linn. Soc.* **61,** 485–499 (1997).
29. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley, New York, 1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Evans, K. Kavanagh, I. Salazar Ciudad, P. Wright, C. Strömberg, A. Gionis, G. Sanson, A. Lister, M. Skinner, I. Pljusnin and J. Eronen for comments and discussions on this work; E. Penttilä for scanning some of the rodents; M. Barbeitos for the suggestion to use information theory; and the following museum curators, collection managers and librarians for loans and reference material: O. Grönwall, R. Asher, M. Hildén, I. Hanski, K. Gully and M. Cytrynbaum. This study was supported by the Academy of Finland (J.J., M.F.), Synthesys (A.R.E.), the Centre for International Mobility (CIMO) (A.R.E.), and a National Science Foundation Postdoctoral Fellowship (G.P.W.).

Author Information Data deposition: the three-dimensional scans for this study are deposited in the MorphoBrowser database (http://morphobrowser.biocenter.helsinki.fi/). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.R.E. (arevans@fastmail.fm).

LETTERS

Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*

Cathy Haag-Liautard[1]*, Mark Dorris[1]*, Xulio Maside[1]†, Steven Macaskill[1]†, Daniel L. Halligan[1], Brian Charlesworth[1] & Peter D. Keightley[1]

Spontaneous mutations are the source of genetic variation required for evolutionary change, and are therefore important for many aspects of evolutionary biology. For example, the divergence between taxa at neutrally evolving sites in the genome is proportional to the per nucleotide mutation rate, *u* (ref. 1), and this can be used to date speciation events by assuming a molecular clock. The overall rate of occurrence of deleterious mutations in the genome each generation (*U*) appears in theories of nucleotide divergence and polymorphism[2], the evolution of sex and recombination[3], and the evolutionary consequences of inbreeding[2]. However, estimates of *U* based on changes in allozymes[4] or DNA sequences[5] and fitness traits are discordant[6–8]. Here we directly estimate *u* in *Drosophila melanogaster* by scanning 20 million bases of DNA from three sets of mutation accumulation lines by using denaturing high-performance liquid chromatography[9]. From 37 mutation events that we detected, we obtained a mean estimate for *u* of 8.4×10^{-9} per generation. Moreover, we detected significant heterogeneity in *u* among the three mutation-accumulation-line genotypes. By multiplying *u* by an estimate of the fraction of mutations that are deleterious in natural populations of *Drosophila*[10], we estimate that *U* is 1.2 per diploid genome. This high rate suggests that selection against deleterious mutations may have a key role in explaining patterns of genetic variation in the genome, and help to maintain recombination and sexual reproduction.

Recurrent deleterious mutations have been implicated in several important evolutionary phenomena. For example, interference between deleterious mutations favours the spread of mutations that increase recombination or sex in finite populations[11]. Synergistic fitness effects of mutations may contribute to the maintenance of recombination and sex in large populations[3]. The positive correlation between recombination rate and nucleotide diversity in several species[12,13] may be caused by linked deleterious mutations reducing diversity in regions of low recombination[2]. Deleterious mutations are also thought to be a major contributor to inbreeding depression[2]. However, the role of deleterious mutations in these and other processes depends on the distribution of fitness effects and the number of deleterious mutations appearing in the genome in each generation (*U*).

Unfortunately, empirical estimates of *U* have been inconsistent and controversial. Two principal methods have been employed to infer *U*. The first is based on differences in fitness traits among mutation accumulation (MA) lines, which are initially genetically uniform and are subsequently maintained at a low population size in benign conditions, so that most new mutations behave neutrally and become fixed at random. However, this method will underestimate *U* because many deleterious mutations are unlikely to affect fitness detectably in the laboratory[6–8]. A second method[14] has no such a bias. *U* is estimated from the product of the mutation rate per nucleotide site per generation (*u*), the number of bases in the diploid genome (2*G*), and the fraction of sites in the genome that are subject to selective constraints (*C*):

$$U = 2uGC \quad (1)$$

C can be estimated from between-species genome comparisons[10,14]. In principle, *u* can be estimated from the nucleotide divergence in unselected genomic regions between a species pair[1] but is subject to uncertainty because the divergence date and generation interval are needed, and identifying neutrally evolving regions can be problematic. Alternatively, *u* can be estimated directly from the molecular divergence between MA lines. The first such estimate was based on electrophoretic mutations in *D. melanogaster*[4], but only three events were detected, and electrophoretic mutations can be related only indirectly to changes in the DNA. More recently, *u* has been estimated in *Caenorhabditis elegans* by sequencing MA-line DNA[5]. From this, an estimate of *U* for coding sequences was obtained, which is one to two orders of magnitude higher than an estimate from the phenotypic divergence of the MA lines, consistent with the expectation outlined above. Here we directly estimate *u* in *D. melanogaster* by scanning the genomes of MA lines, and infer *U* from equation (1).

We scanned 20 megabases (Mb) of DNA, comprising 277 segments (amplicons) of coding, intronic and intergenic DNA (Supplementary Tables S1 and S2, and Supplementary Fig. S1) from 133 MA lines of three genotypes (Florida-33, Florida-39 (ref. 15) and Madrid[16,17]), by denaturing high-performance liquid chromatography (DHPLC)[9]. The efficiency of DHPLC at detecting mutations was verified by analysing synthetic positive controls containing mutations. We successfully detected 45 out of 46 controls (Supplementary Table S3 and Supplementary Fig. S2), which is a similar rate to that in previous reports[18,19]. Putative mutations detected by DHPLC were verified and identified by sequencing. We found evidence for genetic variation in the inbred progenitor of the Florida-39 lines (see Methods and Supplementary Fig. S3 for more details). This manifested itself as fixed nucleotide differences between groups of MA lines for blocks of linked amplicons. Affected amplicons of these lines were excluded from the analysis.

Among 20,002,585 base pairs (bp) screened, we observed 37 mutations (Tables 1 and 2, and Supplementary Fig. S4), of which 3 segregated at a frequency of 0.5 in the line in which they occurred. The mutation detection rate was fairly uniform over the experiment (Supplementary Fig. S5). Our estimate of the single-nucleotide mutation rate per generation is 5.8×10^{-9} (95% confidence interval

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK. †Present addresses: Grupo de Medicina Xenómica, Instituto de Medicina Legal, Universidade de Santiago, S. Francisco s/n, 15782 Santiago de Compostela, Spain (X.M.); Peter MacCallum Cancer Centre, St. Andrews Place, East Melbourne, Victoria 3002, Australia (S.M.).
*These authors contributed equally to this work.

Table 1 | Mutation events detected by DHPLC and confirmed by sequencing

Amplicon	Line	Mutation type	Context
2L-CG8965-C	F33.49	complex cod	CCAAGGA**TG**TCTT→CCAAGGA**CCA**TCTT
3R-113648	M11	complex intron	CATAT**CGTT**CGCA**A**G→CATATCGCA**G**G
X-13003975	M62	complex intron	GATAT(A)$_4$T**T**T**G**CAACTATTTA→GATAT(A)$_4$T**A**T**AT**C**TTA(AT)$_8$T(AT)$_2$(A)$_4$TA**AACTATTT**C**A**
2L-20718966	M87	del.* interg.	GTAGTGTGT**TT...ATGT**AACC→TAAGAGTA(GT)$_3$AACC
2R-CG30377	F33.45	del.† intron	TCTAATG**CG...AG**TCA→TCTAATGTCA
2R-fus	M70	del. cod	AGG(CGG)$_2$**TGG**TTGTG→AGG(CGG)$_2$TTGTG
3R-7922936	F33.67	ins. intron	(T)$_5$AAGG(T)$_9$GTG→(T)$_5$AAGG(T)$_9$**T**GTG
3R-Fru-bis	F33.55	del. intron	AATGACT**CT**GATATT→AATGACTGATATT
3R-19561997	F33.42	del. intron	GGCGT**G**CCAAA→GGCGTCCAAA
X-3198685	M137	del. intron	AGAG(A)$_8$**A**GG→AGAG(A)$_8$GG
X-11335521	M148/M149	ins. intron	TT(A)$_9$CCTTG→TT(A)$_9$**A**CCTTG
3L-22018790	F33.5	TE interg.	CATATGGTAT→CATAT ***(Cr1a)***
2L-cul2-NC2	M78/ M79	ts interg.	AATG**T**ATG→AATG**C**ATG
2L-cul-2-C	F33.8	ts cod	CTT**A**AGCT →CTT**G**AGCT
2R-CG3136	F33.42	ts cod	GCAG**G**TC→GCAG**A**TC
2R-CG30377-up	F39.72	ts interg.	GTCT**T**GAT→GTCT**C**GAT
2R-3097863-down	F33.27	ts interg.	TAAA**C**GGT→TAAA**T**GGT
3L-Bab2-C	F33.8	ts cod	CTGT**G**GGG→CTGT**A**GGG
3L-22018790-down	F33.8	ts interg.	CTAG**G**AAG→CTAG**A**AAG
3L-BcDNAGHO3694	F33.6/ F33.71	ts cod	GGGT**C**ACT→GGGT**T**ACT
3R-113648	F33.49	ts intron	GTCGA**A**GGG→GTCGA**G**GGG
3R-19599719	F39.67	ts intron	ATG**G**GGCG→ATG**A**GGCG
3R-19615776	F39.65	ts cod	ATTTCC**T**TTG→ATTTCC**C**TTG
3R-CG8968	F33.69	ts intron	CATC**G**CTT→CATC**A**CTT
3R-21787667-down	F33.49	ts interg.	CTT**G**CGCT→CTT**A**CGCT
X-11331631-down	M31	ts interg.	GTATA**T**ATGC→GTATA**C**ATGC
X-CG15745	F33.69	ts cod	TGCCC**G**GAG→TGCCC**A**GAG
X-CG15745	M75	ts cod	CGGAA**C**GAG→CGGAA**T**GAG
X-CG32495	F39.67	ts cod	CACC**G**AGG→CACC**A**AGG
2L-CG2955-NC	M73	tv interg.	CAA**T**(T)$_5$AAAG→CAA**A**(T)$_5$AAAG
2L-215156	F33.17/ F33.70	tv interg.	CCGAAA**G**TC→CCGAAA**C**TC
2R-3097863	M137	tv intron	CGAC**T**CAA→CGAC**G**CAA
2R-CG14748	M11	tv cod	GCGG**A**CG→GCGG**T**CG
3L-CG32050	F33.17 /F33.70	tv cod	CACA**A**GAT→CACA**C**GAT
3R-419892	F39.11	tv interg.	GCA**C**AAC→GCA**G**AAC
3R-21787667	M140	tv interg.	GCAT**T**TTGT→GCAT**G**TTGT
X-hiw	M100	tv cod	CAACT**T**GA→CAACT**G**GA

Abbreviations: cod, coding; interg., intergenic; del., deletion; ins., insertion; ts, transition, tv, transversion. *30-bp deletion. †65-bp deletion. Three mutations were segregating at a frequency of 0.5 within their respective MA lines: 2L-CG8965-C, 3L-22018790-down and X-CG32495. Mutations are indicated in bold.

(CI) 2.1×10^{-9} to 1.31×10^{-8}). This is about two-thirds of a direct estimate in *C. elegans*[5]. Our estimate of u for all mutation events is 8.4×10^{-9} (95% CI 3.6×10^{-9} to 1.6×10^{-8}). However, there is significant heterogeneity in u between the three genotypes (likelihood ratio test, $2\log L = 12.5$; $P = 0.002$). In pairwise tests, the mutation rate in Florida-33 is significantly higher than that in Madrid ($2\log L = 12.4$; $P < 0.001$) and nearly significantly higher than in Florida-39 ($2\log L = 3.6$; $P = 0.059$). Transitions were about twice as frequent as transversions (17 *versus* 8, Table 2); this is higher than the roughly 1:1 ratio observed in noncoding polymorphisms in *Drosophila*[20]. Insertion–deletion events (indels) were a minority of the mutations (eight, excluding transposable elements (TEs)). Among these, deletions (six) were more frequent than insertions (two), which is consistent with the high deletion/insertion ratio observed in *Drosophila* pseudogenes[21]. However, our findings are significantly different from the results of sequencing of *C. elegans* MA lines[5], in which indels substantially outnumbered point mutations (Fisher's exact test: $P = 0.05$) and insertions predominated among the indels ($P = 0.02$). Three events involved simultaneous indel and point mutations (Table 1); similar complex events also segregate within some *Drosophila* populations (P. Haddrill, personal communication). We detected only one TE insertion (of the family *Cr1a*), giving an insertion rate per base pair per generation of 2.7×10^{-10} (95% CI 6.8×10^{-12} to 1.5×10^{-9}), corresponding to an insertion rate per diploid of 0.06 per generation (95% CI 0.002 to 0.35). This is not significantly different from estimates obtained by extrapolating movement rates of active TE families in the Madrid MA lines[17]. Mutation rates were similar in coding, intronic and intergenic DNA (Supplementary Table S4; likelihood ratio test of heterogeneity of mutation rates $2\log L = 2.1$, $P = 0.35$), so an effect of transcription-coupled repair is not evident in our data. Two lines had two mutation events (none had more than two), and this is not significantly different from expectation under a Poisson distribution (randomization test: $P > 0.5$).

The euchromatic *Drosophila* genome size, G, is about 118 Mb, so our estimate of the mean diploid genomic mutation rate from all types of mutations is $2uG = 1.99$. From a comparison of the *D. melanogaster* and *D. simulans* genomes, the fraction of point mutations in *Drosophila* that are selectively eliminated, C, is estimated to be 0.58 (ref. 10). From equation (1), assuming that point mutations and indels are equally deleterious on average, the mean genomic deleterious mutation rate is $U = 1.15$ (95% CI 0.49–2.19). However, indels are more likely to be strongly deleterious than point mutations (Supplementary Fig. S6), and including this information gives a slightly higher estimate for U of 1.20 (95% CI 0.51–2.28; Table 2).

We may have underestimated the genomic mutation rate for three reasons. First, hypermutable, repetitive regions are probably underrepresented because amplicons containing them can be difficult to analyse by DHPLC. Second, we may have missed mutations because of the limitations of DHPLC, although our detection rate for positive controls was 98%. Third, C in equation (1) is likely to be an underestimate[10]. If, however, recessive modifiers that increased the mutation rate had become fixed in the MA line progenitors by inbreeding, we might have overestimated U for natural populations. This is a generic problem with experimental estimates of mutation rates that use inbred lines.

Table 2 | Results of scanning the *Drosophila* genome for new mutations

Mutation type	Mutation events detected			
	Madrid	Florida-33	Florida-39	Total
Complex events	2	0.5	0	2.5
Insertions	1	1	0	2
Deletions	3	3	0	6
TEs	0	1	0	1
Transitions	3	9.5	3.5	16
Transversions	5	2	1	8
Total events	14	17	4.5	35.5
Mutation rate parameter	**Mutation rate estimates**			
	Madrid	Florida-33	Florida-39	Overall
u (C + I) $\times 10^9$	2.0 (0.7–4.4)	5.6 (1.9–12.5)	0	2.6 (0.6–9.2)
u (SNM) $\times 10^9$	2.7 (1.2–5.4)	11.7 (5.9–20.6)	6.8 (2.1–16.6)	5.8 (2.1–13.1)
u (total) $\times 10^9$	4.8 (2.6–8.0)	17.2 (10.0–27.6)	6.8 (2.1–16.6)	8.4 (3.6–16.0)
U	0.66 (0.36–1.11)	2.56 (1.49–4.10)	0.94 (0.28–2.28)	1.20 (0.51–2.28)

Totals of 11,207,503 bp, 5,272,760 bp and 3,522,322 bp of Madrid, Florida-33 and Florida-39 DNA were scanned, respectively. u (C + I) is the mutation rate per site for complex and indel events, including TEs. u (SNM) is the mutation rate for single nucleotide mutation events (transitions and transversions). Ranges in parentheses are 95% confidence intervals. The overall estimates of mutation rates are averages, weighted by the average number of lines of each genotype successfully amplified per amplicon. We calculated confidence intervals for the overall mutation rates by maximum likelihood, under the assumption that each genotype's mutation rate is sampled from a log-normal distribution, with Poisson error on mutation numbers within genotypes. We calculated profile likelihoods as a function of the mean of the mutation rate distribution and obtained approximate confidence intervals on the basis of drops of 2 log likelihood units from the maximum likelihoods.

Our findings have several implications. First, we found significant genetic variation in the mutation rate between genotypes. Genetic variation in the mutation rate has been reported in *D. melanogaster*[22], and in the rate by which fitness declines due to MA in rhabditid nematodes[23]. Second, our estimate of the nucleotide site mutation rate is about 5-fold (95% confidence limits 2-fold and 12-fold) higher than a phylogenetic estimate from synonymous site divergence[24], assuming that wild flies undergo ten generations per year. This could be partly due to inaccurate estimates of species divergence times or to differences in generation times between laboratory flies and wild flies. Combined with the recent inference of pervasive selection against new mutations in *Drosophila*[10,25], our estimate for u indicates that U probably exceeds one event per diploid genome per generation in *Drosophila* and is unlikely to be less than 0.5. This is comparable with an estimate in *C. elegans* based on direct sequencing (0.96 for coding sequences only[5]).

However, genomic deleterious mutation rates estimated from the divergence of fitness traits in MA lines strongly disagree between these species; these are about 0.01 in *C. elegans*[23] and up to about 1.0 in *Drosophila*[6–8,26]. The distribution of fitness effects of deleterious mutations in *Drosophila* is likely to be highly leptokurtic[27], so it is unexpected that some *Drosophila* MA experiments should yield similar phenotypic[6] and DNA-based (our study) estimates of U. The reasons for this discrepancy remain obscure[2,7,8,23,26]. Last, our results have implications for the evolutionary maintenance of sex and recombination. Non-zero rates of recombination can be maintained by both Hill–Robertson interference[11] and synergistic epistasis[28], with genomic deleterious mutation rates as low as 0.5 (our lower confidence limit). However, our estimate of $U = 1.2$ seems too low for deterministic selection against deleterious mutations to allow the maintenance of sexual reproduction with a twofold cost, although the mechanism might work if U were as high as our upper confidence limit[28]. Additional factors that slow the spread of asexual mutants, such as population structure[29], might help to maintain sex in species with suitable population structure, even with U as low as 0.5.

METHODS

Mutation accumulation lines. We analysed *D. melanogaster* MA lines of three genotypes (Florida-33, Florida-39 and Madrid). Progenitors of Florida-33 and Florida-39 were derived independently from a common base population by brother–sister mating for 40 generations, then MA lines were maintained by full-sib mating until generation 90 (ref. 15), and by a mixture of full-sib and half-sib mating until generation 187, on average (D. Houle, personal communication). The Madrid progenitor was established by chromosome extraction[16]. MA lines were maintained by full-sib or double first-cousin mating until generation 47, then by full-sib mating until generation 262 (refs 16, 17). DNA was extracted from pools of 25 individuals per line.

Mutation detection by DHPLC. We randomly selected 77 nucleotide positions from the euchromatic genome sequence of *D. melanogaster* (Release 4.3 for chromosome 4, otherwise Release 3.1). A coding and a non-coding amplicon, each of 650–750 bp, were chosen close to each position. At an additional 56 random positions we selected either a coding or a non-coding amplicon. Finally, we selected 67 non-coding amplicons flanking suspected mutations (see below). For each amplicon, 5 ng of template was amplified by PCR with AmpliTaq Gold (Applied Biosystems), and the length and quality of products were verified on 1% agarose gels. Significantly weaker products than the others were excluded, because detection of variants at a frequency of less than 10% in a pooled sample is unreliable. The sequences of PCR products of the same MA-line genotype were compared by DHPLC[9]. Products were mixed in groups of four (labelled 'vials'), the mixtures were denatured and reannealed, and the fragments were separated on a Transgenomic Wave 3500A DHPLC instrument with a DNASep column at two to five temperatures with elution gradients chosen according to the sequence of the amplicon. In the absence of a mutation, vials gave similar elution profiles. If a line carried a mutation, the difference in retention time between heteroduplexes and homoduplexes resulted in its vial showing a wider profile or a double peak (Supplementary Fig. S4).

Positive controls. We assessed the detection rate of mutations by using positive controls for 46 amplicons, generated with the use of the relatively high misincorporation rate of traditional *Taq* polymerase. We amplified MA-line genomic DNA with a non-proofreading polymerase, and cloned and sequenced PCR products. Positive and negative controls were selected among the clones with one and zero mutations, respectively, compared with the wild-type sequence (Supplementary Table S3). Controls were then amplified by PCR along with the MA lines. Mixtures were produced between the positive control product and products from three MA lines of the same genotype, the negative and positive control products, and products of the negative control and the three MA lines of the same genotype. These were analysed by DHPLC along with the MA-line vials of that amplicon. Most positive controls are transitions (Supplementary Table S3), which are more difficult to detect by DHPLC than indels or transversions[30], making our positive control panel conservative.

Characterization of mutations. Whenever DHPLC elution profiles showed differences, the four lines of that vial were reamplified by PCR and directly sequenced in both directions. Some mutations were found to be segregating within a line; the strategy we used to investigate these is described in Supplementary Table S4.

Polymorphic sites not representing new mutations. Polymorphisms present at the start of the MA phase are expected to become fixed in different lines, and polymorphism is likely to affect a chromosomal region. We detected several regions having such characteristics in Florida-39, but not in Florida-33 or Madrid (Supplementary Fig. S3). Lines showing polymorphism in Florida-39 were excluded from the data on affected amplicons. Furthermore, to distinguish between genuine mutations and polymorphism blocks, noncoding amplicons closely linked to either side of putative mutations were analysed. This procedure makes it improbable that a polymorphism would be misclassified as a mutation

(Supplementary Fig. S3). In four cases, pairs of lines shared identical mutations (Table 1). In particular, the Madrid lines involved were consecutively numbered, and one of the Florida-33 events concerned lines sharing two mutations. These events presumably reflect breeding contamination between two MA lines[17]. We counted shared mutations only once, and reduced the total number of lines by 0.5 for each contaminant.

Received 9 August; accepted 30 October 2006.

1. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
2. Charlesworth, B. & Charlesworth, D. Some evolutionary consequences of deleterious mutations. *Genetica* **102–103,** 3–19 (1998).
3. Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336,** 435–440 (1988).
4. Mukai, T. & Cockerham, C. C. Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster. Proc. Natl Acad. Sci. USA* **74,** 2514–2517 (1977).
5. Denver, D. R., Morris, K., Lynch, M. & Thomas, W. K. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430,** 679–682 (2004).
6. Crow, J. F. & Simmons, M. J. in *The Genetics and Biology of Drosophila* Vol. 3C (eds Ashburner, M., Carson, H. L. & Thompson, J. N.) 1–35 (Academic, London, 1983).
7. Keightley, P. D. & Eyre-Walker, A. Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153,** 515–523 (1999).
8. Lynch, M. *et al.* Perspective: Spontaneous deleterious mutation. *Evolution* **53,** 645–663 (1999).
9. Oefner, P. J. & Huber, C. G. A decade of high-resolution liquid chromatography of nucleic acids on styrene divinylbenzene copolymers. *J. Chromatogr. B* **782,** 27–55 (2002).
10. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16,** 875–884 (2006).
11. Keightley, P. D. & Otto, S. P. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* **443,** 89–92 (2006).
12. Presgraves, D. C. Recombination enhances protein adaptation in *Drosophila melanogaster. Curr. Biol.* **15,** 1651–1656 (2005).
13. Nachman, M. W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17,** 481–485 (2001).
14. Kondrashov, A. S. & Crow, J. F. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2,** 229–234 (1993).
15. Houle, D. & Nuzhdin, S. V. Mutation accumulation and the effect of *copia* insertions in *Drosophila melanogaster. Genet. Res.* **83,** 7–18 (2004).
16. Fernandez, J. & López-Fanjul, C. Spontaneous mutational variances and covariances for fitness-related traits in *Drosophila melanogaster. Genetics* **143,** 829–837 (1996).
17. Maside, X., Bartolome, C., Assimacopoulos, S. & Charlesworth, B. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: *in situ* hybridization *vs* Southern blotting data. *Genet. Res.* **78,** 121–136 (2001).
18. Dobson-Stone, C. *et al.* Comparison of fluorescent single-strand conformation polymorphism analysis and denaturing high-performance liquid chromatography for detection of EXT1 and EXT2 mutations in hereditary multiple exostoses. *Eur. J. Hum. Genet.* **8,** 24–32 (2000).
19. O'Donovan, M. C. *et al.* Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics* **52,** 44–49 (1998).
20. Moriyama, E. N. & Powell, J. R. Intraspecific nuclear DNA variation in *Drosophila. Mol. Biol. Evol.* **13,** 261–277 (1996).
21. Petrov, D. A. DNA loss and evolution of genome size in *Drosophila. Genetica* **115,** 81–91 (2002).
22. Woodruff, R. C., Thompson, J. N., Seeger, M. A. & Spivey, W. E. Variation in spontaneous mutation and repair in natural population lines of *Drosophila melanogaster. Heredity* **53,** 223–234 (1984).
23. Baer, C. F. *et al.* Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. *Proc. Natl Acad. Sci. USA* **102,** 5785–5790 (2005).
24. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21,** 36–44 (2004).
25. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila. Nature* **437,** 1149–1152 (2005).
26. Fry, J. D. On the rate and linearity of viability declines in *Drosophila* mutation-accumulation experiments: Genomic mutation rates and synergistic epistasis revisited. *Genetics* **166,** 797–806 (2004).
27. Loewe, L. & Charlesworth, B. Inferring the distribution of mutational effects on fitness in *Drosophila. Biol. Lett.* **2,** 426–430 (2006).
28. Charlesworth, B. Mutation selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.* **55,** 199–221 (1990).
29. Salathé, M., Schmid-Hempel, P. & Bonhoeffer, S. Mutation accumulation in space and the maintenance of sexual reproduction. *Ecol. Lett.* **9,** 941–946 (2006).
30. Ravnik-Glavač, M., Atkinson, A., Glavač, D. & Dean, M. DHPLC screening of cystic fibrosis gene mutations. *Hum. Mutat.* **19,** 374–383 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Houle and C. López-Fanjul for providing samples of MA lines, P. Andolfatto for suggesting the use of PCR errors as positive controls, F. Oliver for help with DNA sequencing, and D. Charlesworth, J. Crow, J. Drake, A. Eyre-Walker, C. Haag, D. Houle and M. Lynch for comments on the manuscript. We are grateful to the Wellcome Trust for funding by a Functional Genomics Development Initiative grant.

Author Contributions S.M., C.H.-L. and M.D. performed the DHPLC analysis. M.D. cloned and sequenced putative variants. X.M. cloned and sequenced positive controls. D.L.H. analysed selective constraints on indel mutations. B.C. advised on *Drosophila* genetics and interpreting the data. P.D.K. conceived and designed the project. C.H.-L. and P.D.K. analysed the data and wrote the paper. All authors revised the draft manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.D.K. (keightley.drosmutrate@spambob.net).

CORRIGENDUM

doi:10.1038/nature06946

Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*

Cathy Haag-Liautard, Mark Dorris, Xulio Maside, Steven Macaskill, Daniel L. Halligan, David Houle[1], Brian Charlesworth & Peter D. Keightley

[1]Department of Biological Science, Florida State University, Tallahassee, Florida 32306-1100, USA.

Nature 445, 82–85 (2007)

In this Letter, David Houle was omitted from the author list. David Houle was responsible for producing the Florida mutation accumulation lines that were analysed in the experiment.

LETTERS

Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*

Walton D. Jones[1], Pelin Cayirlioglu[2], Ilona Grunwald Kadow[2]† & Leslie B. Vosshall[1]

Blood-feeding insects, including the malaria mosquito *Anopheles gambiae*, use highly specialized and sensitive olfactory systems to locate their hosts. This is accomplished by detecting and following plumes of volatile host emissions, which include carbon dioxide (CO_2)[1]. CO_2 is sensed by a population of olfactory sensory neurons in the maxillary palps of mosquitoes[2,3] and in the antennae of the more genetically tractable fruitfly, *Drosophila melanogaster*[4]. The molecular identity of the chemosensory CO_2 receptor, however, remains unknown. Here we report that CO_2-responsive neurons in *Drosophila* co-express a pair of chemosensory receptors, *Gr21a* and *Gr63a*, at both larval and adult life stages. We identify mosquito homologues of *Gr21a* and *Gr63a*, *GPRGR22* and *GPRGR24*, and show that these are co-expressed in *A. gambiae* maxillary palps. We show that *Gr21a* and *Gr63a* together are sufficient for olfactory CO_2-chemosensation in *Drosophila*. Ectopic expression of *Gr21a* and *Gr63a* together confers CO_2 sensitivity on CO_2-insensitive olfactory neurons, but neither gustatory receptor alone has this function. Mutant flies lacking *Gr63a* lose both electrophysiological and behavioural responses to CO_2. Knowledge of the molecular identity of the insect olfactory CO_2 receptors may spur the development of novel mosquito control strategies designed to take advantage of this unique and critical olfactory pathway. This in turn could bolster the worldwide fight against malaria and other insect-borne diseases.

Carbon dioxide is a pervasive chemical stimulus that is important in the ecology of many insect species[5]. Interestingly, the ethological message conveyed by this gas is highly species- and context-specific. The hawkmoth, *Manduca sexta*, evaluates the quality of *Datura wrightii* flowers by measuring the amount of CO_2 that a given flower produces[6]; newly opened flowers emit more CO_2 and are preferred because they offer more nectar. In response to elevated CO_2 in their hives, honeybees show a stereotyped fanning response that ventilates the hive and reduces ambient CO_2 levels[7]. For blood-feeding female mosquitoes, CO_2 emitted in the breath of animal hosts (~4–5%) is an arousing stimulus that synergizes with host body odour to produce host-seeking behaviours[1,8]. The ecological relevance of CO_2 to fruitflies is less clear, but CO_2 is one component of the aversive *Drosophila* stress odorant (dSO)[9] and may also signal food source suitability[10].

The chemosensory neurons that are thought to underlie these CO_2-evoked behaviours have been functionally characterized on antennal, maxillary or labial appendages in a number of different insects[11]. *Drosophila* antennae have a small, CO_2-sensitive subpopulation of olfactory sensory neurons that have been designated ab1C (ref. 4). The antennal lobe of the fly brain has a single ventrally-situated glomerulus (V) that responds selectively to CO_2 (ref. 9). This V glomerulus is innervated by a population of olfactory neurons expressing the chemosensory receptor *Gr21a* (ref. 12); these neurons correspond to the ab1C neurons. *Gr21a* is a member of the gustatory receptor gene family, which includes bitter and sweet taste receptors necessary for taste recognition in the fly, along with a number of gustatory receptor genes expressed in the antenna that may function as odorant receptors[12]. Although clearly separable into two gene families, *Drosophila* gustatory receptors and odorant receptors are thought to have a common phylogenetic origin and were originally assigned to taste and smell modalities by gene homology and not function[13]. Genetic silencing[9] or ablation[10] of *Gr21a*-expressing neurons eliminates both adult[9] and larval[10] chemosensory responses to CO_2, confirming that these are the only CO_2-sensitive neurons in *Drosophila*.

We asked whether *Gr21a* is merely a marker for the CO_2-sensitive neurons in *Drosophila* or whether it is directly involved in CO_2 detection. As it has been previously reported that taste neurons can express multiple gustatory receptor genes[14], we began by screening for additional gustatory receptor genes expressed in ab1C neurons. Two other gustatory receptor genes are known to be expressed in the antenna[12], and fluorescent RNA *in situ* hybridization reveals that *Gr63a* is co-expressed with *Gr21a* (Fig. 1a), but that *Gr10a* is expressed in the adjacent ab1D neuron (Fig. 1b)[15]. Confirming these *in situ* hybridization results, neurons labelled with genetic markers under the control of *Gr21a* and *Gr63a* promoters co-converge upon the CO_2-sensitive V glomerulus (Fig. 1c). These chemosensory receptors are therefore co-expressed in the adult ab1C sensillum (Fig. 1d).

Next, we investigated the expression of *Gr21a* and *Gr63a* in the larval olfactory system. Larvae show robust avoidance of CO_2, which is mediated by *Gr21a*-expressing neurons[10]. Both *Gr21a-GAL4* and *Gr63a-GAL4* transgenes drive expression of a membrane-tethered green fluorescent protein (GFP) in the same neuron that innervates the larval terminal organ, which is thought to be primarily gustatory in function (Fig. 1e). This indicates that *Gr21a* and *Gr63a* are also co-expressed in the larval chemosensory system.

To generalize our results to other insects, we analysed *Gr21a* and *Gr63a* homologues in *A. gambiae*. Mosquitoes, in whom CO_2 plays an important role in human-host-seeking, have closely related homologues of both *Gr21a* and *Gr63a*, called *GPRGR22* and *GPRGR24*, respectively[16] (Fig. 1f). RNA *in situ* hybridization reveals co-expression of *GPRGR22* and *GPRGR24* in a subset of neurons in the maxillary palp, the CO_2-sensitive organ of the mosquito (Fig. 1g). No expression is detected in the antenna or proboscis (data not shown). As in the fly, these putative mosquito CO_2-responsive neurons do not express *GPROR7*, the *A. gambiae Or83b* orthologue (Fig. 1h). Thus these mosquito homologues share three key properties in common with fly *Gr21a*/*Gr63a*: they are co-expressed in the same sensory neurons; they are selectively expressed only in the olfactory appendage that responds to CO_2; and they do not express the olfactory co-receptor *Or83b*.

[1]Laboratory of Neurogenetics and Behavior, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. [2]Department of Biological Chemistry, Howard Hughes Medical Institute, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1662, USA. †Present address: Department of Molecular Neurobiology, Max-Planck Institute of Neurobiology, Am Klopferspitz 18, 82152 Martinsried, Germany.

To investigate the role of these gustatory receptor genes as putative CO_2 receptors, we ectopically expressed the antennal gustatory receptor genes both alone and in pairs in neurons normally unresponsive to CO_2 using the *GAL4*/UAS system (Fig. 2a). *Or22a-GAL4* drives expression in ~75% of the electrophysiologically accessible ab3A neurons that express *Or22a/b* and the co-receptor *Or83b* (ref. 17). No individual gustatory receptor gene was capable of conferring CO_2 responsiveness on the ab3A neurons (Fig. 2b), but as we previously demonstrated that fly odorant receptor genes are obligate OR/Or83b heterodimers[18], we asked whether a combination of two gustatory receptor genes could function as a CO_2 receptor. Neither *Gr21a* nor *Gr63a* confer responses to CO_2 when combined with *Gr10a*, but the combination of *Gr21a* and *Gr63a* produces a significant response to a stimulus of ~3% CO_2 (Fig. 2b). It is therefore the specific combination of these two gustatory receptor genes that is sufficient to induce CO_2 sensitivity rather than a generic requirement for the co-expression of any two antennal gustatory receptor genes. *Gr21a* and *Gr63a* together also increase the level of spontaneous activity in the ab3A neuron. We considered the possibility that this reflects activity in response to ambient CO_2 levels (0.035%), but we found that the activity of these neurons is not reduced in response to a CO_2-free air stream (data not shown). Prior results with odorant receptor genes indicate that some have substantial odour-independent activity[19], and this result suggests that gustatory receptor genes share this property. Further analysis of ectopically expressed *Gr21a*/*Gr63a* reveals a dose-dependent increase to stimuli of increasing CO_2 concentration, whereas animals expressing *Gr21a* alone do not respond to CO_2 at any concentration tested (Fig. 2c, compare blue and green curves).

We next compared the efficacy of the ectopic CO_2 response to that obtained in the native ab1C sensillum, and find that although the dose–response curves have a similar shape, the efficacy of the ectopic receptor is lower than the endogenous response (Fig. 2c, compare red and blue curves). A number of explanations may account for this difference in efficacy. There could be a requirement for a cell-type specific co-factor, which is missing in the ab3A cell. Alternatively, lower receptor expression levels or competition for trafficking factors with the resident odorant receptors in ab3A could lead to a reduced

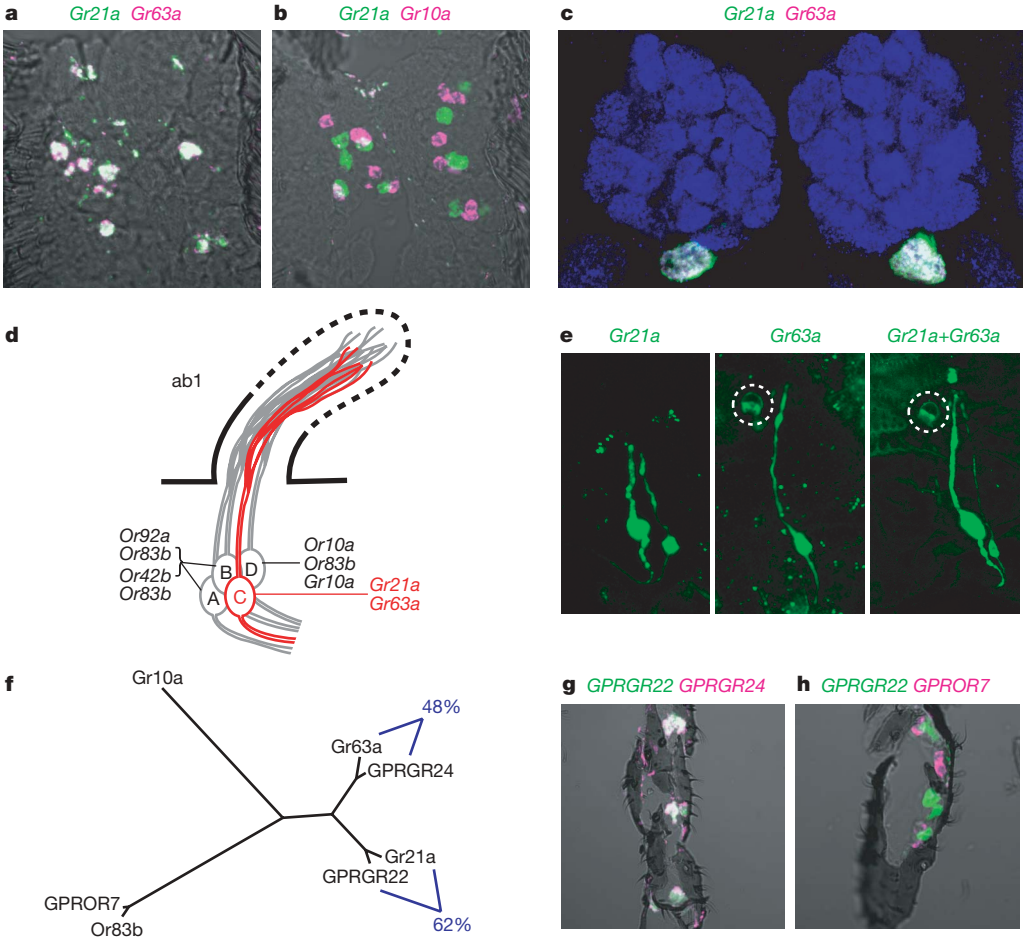


Figure 1 | *Gr21a* and *Gr63a* are co-expressed in the CO_2-responsive chemosensory neurons. **a**, Fluorescent double *in situ* hybridization on the third antennal segment of wild-type *Drosophila* (wild-type Berlin) reveals co-expression of *Gr21a* (green) and *Gr63a* (magenta) mRNA. **b**, *Gr21a* mRNA (green) is not co-expressed with the only other gustatory receptor gene expressed in the antenna, *Gr10a* (magenta). **c**, Olfactory neurons expressing *Gr21a-GAL4*; UAS-*CD8-GFP* (green) and *Gr63a-sytRFP* (magenta) co-converge upon the V glomerulus in the antennal lobe. Whole mount brain immunofluorescence preparation is counter-stained with the neuropil marker nc82 (blue)[28]. **d**, Diagram of the ab1 sensillum, with the CO_2-responsive ab1C neuron labelled in red. The receptor genes expressed in ab1D are indicated. The receptor pairs expressed in the two remaining neurons have not been conclusively assigned to either ab1A or ab1B. **e**, Larvae expressing a membrane-tethered GFP under control of *Gr21a-GAL4* (left) have two labelled neurons innervating the terminal organ, one of which probably represents ectopic *GAL4* expression as seen in other larval receptor *GAL4* lines[29]. This smaller cell of unknown function is visible in both the left and right panels. *Gr63a-GAL4* (middle) labels one neuron, and the combination of *Gr21a-GAL4* and *Gr63a-GAL4* (right) labels the same two neurons as *Gr21a-GAL4* alone, indicating co-expression. When visible, the adjacent olfactory dorsal organ is encircled by a white dotted line. **f**, Phylogenetic comparison of *Drosophila* CO_2 receptors and their nearest *A. gambiae* homologues, with percentage amino acid identity in blue. **g**, RNA *in situ* hybridization of *A. gambiae* maxillary palps reveals co-expression *GPRGR22* (green) and *GPRGR24* (magenta). **h**, mRNAs for *GPRGR22* (green) and *GPOR7* (magenta) are not co-expressed.

number of functional CO_2 receptors. Such competition, leading to lower efficacy of ectopically expressed chemosensory receptors, has been noted by other investigators[17,18]. Finally, there is evidence that the ab1C neuron has a uniquely specialized dendritic architecture, with considerably more branching than other chemosensory neurons[20]. These special structural properties may be necessary for optimal Gr21a/Gr63a receptor function and may constrain its efficacy in other neurons. Taken together, we find that *Gr21a* and *Gr63a* together are sufficient to confer dose-dependent CO_2-responsivity on olfactory neurons normally unresponsive to CO_2.

To investigate the role of these gustatory receptor genes in the CO_2 responses of the native ab1C neuron, we screened for *Gr21a* and *Gr63a* null mutants by homologous recombination (Fig. 3a)[21,22]. *Gr21a* proved to be resistant to mutagenesis, but we obtained a single null mutant allele of *Gr63a*. PCR (polymerase chain reaction) analysis of $Gr63a^1$ indicates the selective loss of *Gr63a* without affecting a neighbouring gene, *CG1079* (Fig. 3a). $Gr63a^1$ flies lack the *Gr63a* transcript when compared with parental controls, but have normal levels of *Gr21a* (Fig. 3a). Electrophysiological recordings of ab1 sensilla in $Gr63a^1$ flies reveal a complete indifference to stimuli of ~2.25% CO_2, in stark contrast to wild-type parental control flies, whose ab1C neurons respond strongly. The $Gr63a^1$ allele is genetically recessive, because the sensilla of heterozygous individuals have an essentially wild-type CO_2 response. CO_2 responses in the $Gr63a^1$ are restored by rescuing *Gr63a* expression in the ab1C neurons using the *GAL4*/UAS system, while control $Gr63a^1$ flies bearing either the *Gr21a-GAL4* transgene or the UAS-*Gr63a* transgene alone fail to respond (Fig. 3b).

As genetic silencing of *Gr21a*-expressing neurons eliminates olfactory CO_2 avoidance in a T-maze[9], we asked whether $Gr63a^1$ flies have CO_2 avoidance defects. Whereas wild-type flies robustly avoid CO_2 in a T-maze, $Gr63a^1$ flies fail to distinguish room air from a ~2% CO_2 stimulus. Consistent with their electrophysiological responses, $Gr63a^1$ heterozygotes show a wild-type avoidance response, whereas $Gr63a^1$ flies bearing either *Gr21a-GAL4* or UAS-*Gr63a* transgenes fail to differentiate room air from 2% CO_2. When combined, however, these two transgenes rescue olfactory CO_2 avoidance in the mutant (Fig. 4). The failure of the rescue to reach wild-type levels in either the electrophysiological recordings or the behaviour is probably a consequence of the lower levels of *Gr63a* expression in rescued ab1C neurons when compared to wild-type ab1C neurons, as also discussed above (data not shown). These loss of function results prove that *Gr63a* is necessary for CO_2 chemoreception in *Drosophila* and strengthen our hypothesis that the *Drosophila* CO_2 receptor is composed of both Gr21a and Gr63a.

Taken together, our results suggest that two chemosensory receptors, *Gr21a* and *Gr63a*, are necessary and sufficient for detection of CO_2 in *Drosophila*. Despite the fact both *Gr21a* and *Gr63a* are required for CO_2 responses, our data at present do not allow us to resolve whether one subunit acts as a chaperoning co-receptor while the other subunit confers ligand specificity (as is the case for odorant receptors and Or83b), or whether both subunits are required for both functions. This is because, unfortunately, our attempts to tag these proteins while retaining function have failed. Previous work in other biological systems has implicated several cytosolic proteins as gas sensors. Atypical soluble guanylate cyclases are candidate oxygen sensors in *Caenorhabditis elegans*[23], while conventional soluble guanylate cyclases are cytosolic receptors for nitric oxide and carbon monoxide[24,25]. A nuclear receptor in *Drosophila* has been suggested as an additional receptor for nitric oxide and carbon monoxide[26]. Bacteria utilize haem-containing myoglobin in chemotaxis towards oxygen[27]. Although our genetic evidence strongly suggests that *Gr21a*/*Gr63a* encodes the first example of a membrane-associated gas sensor, we cannot exclude the possibility that additional secreted or cytosolic proteins, such as those described previously, are essential co-factors for CO_2 detection. Further biochemical investigation—most effectively carried out in a cell-based heterologous expression system—will be required to address this question.

It remains to be resolved whether the Gr21a/Gr63a receptor binds gaseous CO_2 or a metabolite, such as bicarbonate (HCO_3^-). It will also be of interest to elucidate the signal transduction cascade to which these gustatory receptors couple in order to transform the absolute concentration of environmental CO_2 into precise trains of neuronal action potentials. As CO_2 is an important stimulus for a large number of insect pests, the identification of the CO_2 receptor provides a potential target for the design of inhibitors that might be useful as insect repellents. These would be important weapons in the fight against global infectious disease by reducing the attraction of blood-feeding insects to human hosts.

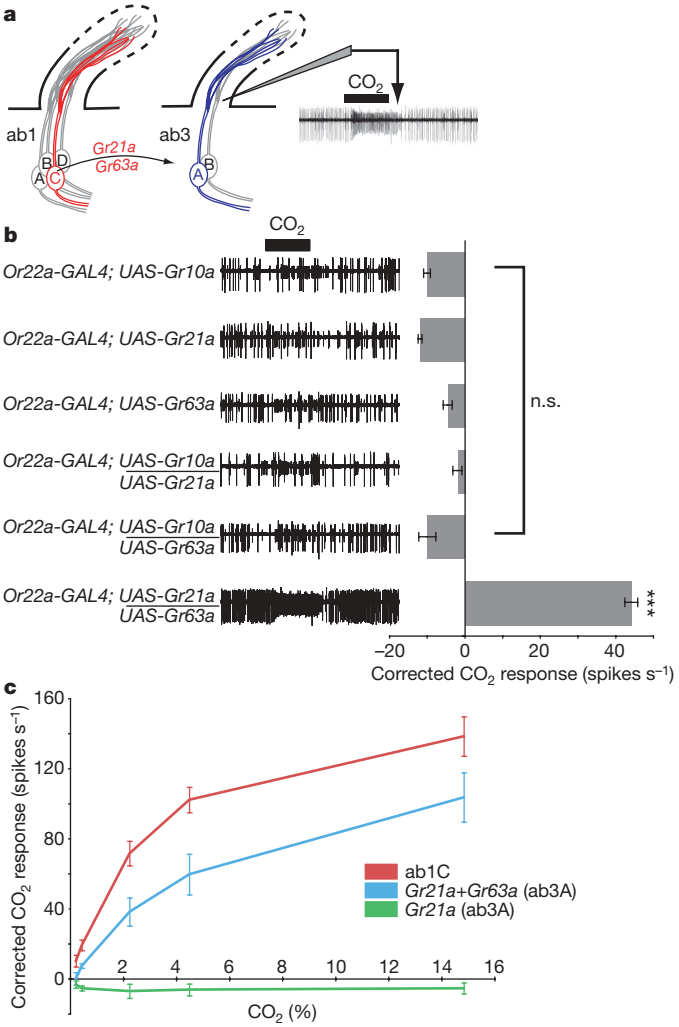


Figure 2 | Expression of both *Gr21a* and *Gr63a* confers CO_2 sensitivity on normally CO_2-insensitive neurons. **a**, Diagram outlining the methodology for the transfer of the CO_2 receptor to the ab3A neuron. Extracellular recording of spikes emitted by CO_2 is at the right. **b**, The indicated combinations of antennal gustatory receptor genes were ectopically expressed in ab3A neurons using the *Or22a-GAL4* driver. Single-sensillum electrophysiological recordings on transgenic ab3 sensilla, recognized by their characteristic response to ethyl hexanoate (ab3A) and 2-heptanone (ab3B)[17], were made for both room air (~0.035% CO_2) and ~3% CO_2. Representative traces (stimulus bar, 1 s) are on the left, and mean responses (±s.e.m.) are on the right (n = 15–18 sensilla per genotype). Significant responses to CO_2 are only found with the combination of *Gr21a* and *Gr63a* (Tukey HSD test; $P < 10^{-6}$). n.s., not significant. **c**, Dose–response curves for the combination of *Gr21a* and *Gr63a* (blue curve) versus expression of *Gr21a* alone (green curve) in ab3A neurons as compared to the CO_2 response in native ab1C neurons (red curve). Mean responses (s.e.m.) are plotted (n = 10 sensilla per genotype).

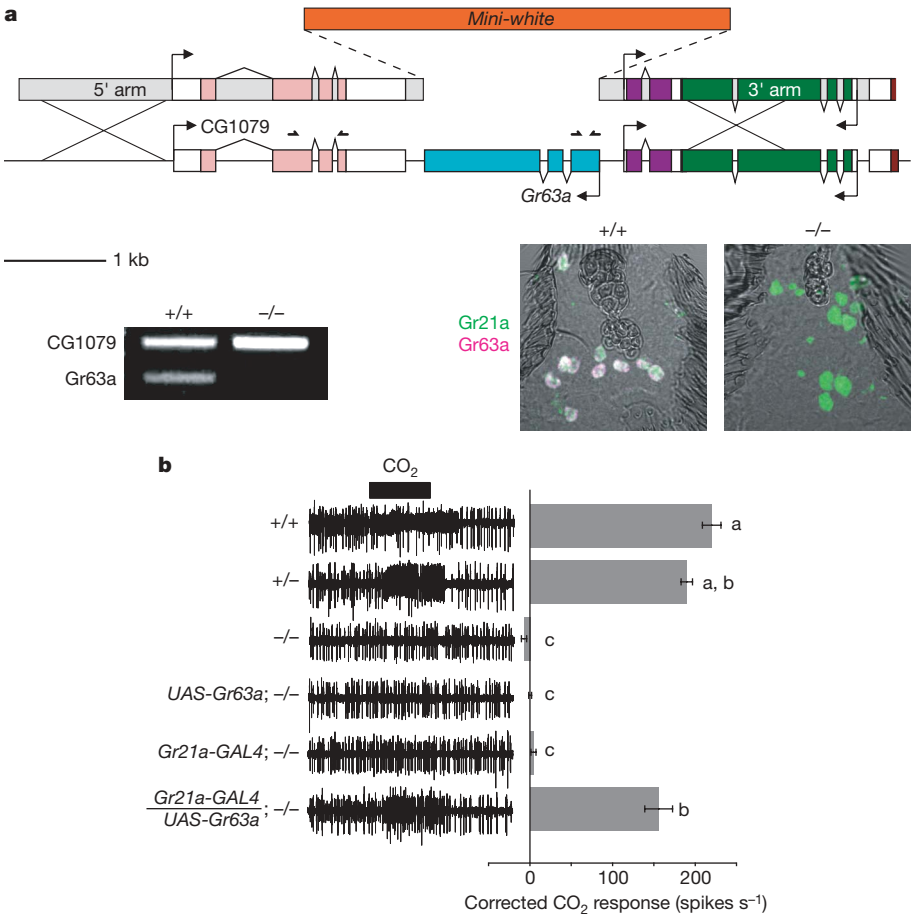


Figure 3 | *Gr63a*^{1} mutants are electrophysiologically and behaviourally insensitive to CO_2. **a**, Diagram of the *Gr63a* gene targeting construct and the *Gr63a* genomic locus. PCR using primers denoted by small arrows above the locus diagram indicates the absence of *Gr63a* product in *Gr63a*^{1} mutants. *Gr63a*- flies lack *Gr63a* mRNA (magenta) compared to wild-type controls, while *Gr21a* mRNA (green) levels remain unchanged. **b**, *Gr63a*^{1} mutant ab1 sensilla (−/−) do not respond to ~2.25% CO_2 when compared to parental wild-type or heterozygous (+/−) flies. Responses are rescued by the combination of *Gr21a-GAL4* and UAS-*Gr63a* in the mutant *Gr63a*^{1} background but not by either transgene alone. Representative traces (stimulus bar, 1 s) at left and mean responses (±s.e.m.; $n = 12$ for all genotypes) on the right were quantified as in Fig. 2. Statistical significance was calculated using a Tukey HSD test comparing all pairs of means ($P < 0.001$). Bars labelled with different letters are significantly different.

METHODS

RNA *in situ* hybridization and immunofluorescence. Double fluorescent RNA *in situ* hybridization was performed on fly antennae as described previously[15], and on mosquito maxillary palps without protocol modification. Adult mosquitoes (*A. gambiae* strain G3; MRA-112) were obtained from MR4. Whole mount brain and larval immunostaining was performed as previously described[28,29]. Details can be found in Supplementary Methods.

Gustatory receptor transgene generation. *Gr10a* was amplified from Oregon-R antennal complementary DNA, and *Gr63a* was amplified from *yw* genomic DNA. *GPRGR22* and *GPRGR24* were amplified from *A. gambiae* G3 antennal cDNA. Construct details can be found in Supplementary Methods.

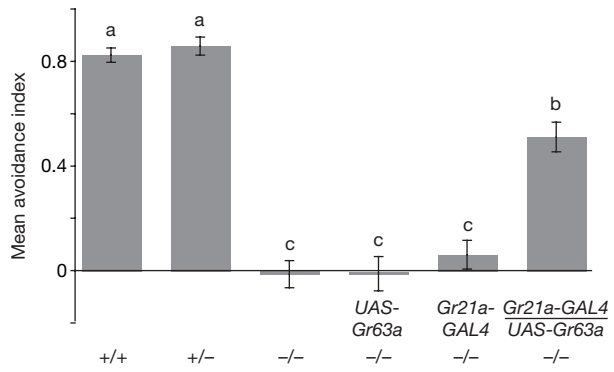


Figure 4 | *Gr63a*^{1} mutants and the *GAL4* and UAS controls are all indifferent to CO_2 in a T-maze, whereas wild-type and heterozygous *Gr63a*^{1} flies show robust avoidance. This deficit is rescued in *Gr21a-GAL4*/UAS-Gr63a; *Gr63a*^{1} flies. Mean avoidance ± s.e.m. is indicated ($n = 15$). Statistical significance was calculated using a Tukey HSD test comparing all pairs of means ($P < 0.01$). Bars labelled with different letters are significantly different.

Single sensillum electrophysiology. Extracellular recordings of ab1 and ab3 sensilla from individual flies (2–10 days old) were made as described[4,22] and as in the Supplementary Methods.

***Gr63a* targeting construct and mutant screen.** Genomic DNA both 5′ and 3′ of the *Gr63a* coding sequence was amplified from *yw* flies and cloned into the CMC105 gene targeting vector[22] (Supplementary Methods). Four independent insertions of the targeting construct were screened as described[22]. The progeny of approximately 16,500 virgin mosaic or white-eyed females (~330,000 flies) were screened for re-insertion on the third chromosome, and we recovered a single mutant allele, *Gr63a*^{1}. PCR confirmation of *Gr63a*^{1} was performed on genomic DNA preparations of the mutant line and its corresponding wild-type parental targeting construct insertion with primers within *Gr63a* itself and within the neighbouring gene CG1079 (Supplementary Methods). A similar screen for a *Gr21a* mutant produced no mutants among ~350,000 progeny derived from five independent targeting construct insertions.

Received 31 October; accepted 23 November 2006.
Published online 13 December 2006.

1. Gillies, M. T. The role of carbon dioxide in host-finding in mosquitoes (Diptera:Culicidae): a review. *Bull. Entomol. Res.* **70**, 525–532 (1980).
2. Kellogg, F. E. Water vapour and carbon dioxide receptors in *Aedes aegypti*. *J. Insect Physiol.* **16**, 99–108 (1970).
3. Grant, A. J., Wigton, B. E., Aghajanian, J. G. & O'Connell, R. J. Electrophysiological responses of receptor neurons in mosquito maxillary palp sensilla to carbon dioxide. *J. Comp. Physiol. A* **177**, 389–396 (1995).
4. de Bruyne, M., Foster, K. & Carlson, J. R. Odor coding in the *Drosophila* antenna. *Neuron* **30**, 537–552 (2001).
5. Nicolas, G. & Sillans, D. Immediate and latent effects of carbon dioxide on insects. *Annu. Rev. Entomol.* **34**, 97–116 (1989).
6. Thom, C., Guerenstein, P. G., Mechaber, W. L. & Hildebrand, J. G. Floral CO_2 reveals flower profitability to moths. *J. Chem. Ecol.* **30**, 1285–1288 (2004).
7. Southwick, E. E. & Moritz, R. F. A. Social control of air ventilation in colonies of honey bees, *Apis mellifera*. *J. Insect Physiol.* **33**, 623–626 (1987).
8. Takken, W. & Knols, B. G. Odor-mediated behavior of Afrotropical malaria mosquitoes. *Annu. Rev. Entomol.* **44**, 131–157 (1999).
9. Suh, G. S. *et al.* A single population of olfactory sensory neurons mediates an innate avoidance behaviour in *Drosophila*. *Nature* **431**, 854–859 (2004).

10. Faucher, C., Forstreuter, M., Hilker, M. & de Bruyne, M. Behavioral responses of *Drosophila* to biogenic levels of carbon dioxide depend on life-stage, sex and olfactory context. *J. Exp. Biol.* **209,** 2739–2748 (2006).
11. Stange, G. & Stowe, S. Carbon-dioxide sensing structures in terrestrial arthropods. *Microsc. Res. Tech* **47,** 416–427 (1999).
12. Scott, K. *et al.* A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila. Cell* **104,** 661–673 (2001).
13. Robertson, H. M., Warr, C. G. & Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster. Proc. Natl Acad. Sci. USA* **100** (Suppl. 2), 14537–14542 (2003).
14. Wang, Z., Singhvi, A., Kong, P. & Scott, K. Taste representations in the *Drosophila* brain. *Cell* **117,** 981–991 (2004).
15. Fishilevich, E. & Vosshall, L. B. Genetic and functional subdivision of the *Drosophila* antennal lobe. *Curr. Biol.* **15,** 1548–1553 (2005).
16. Hill, C. A. *et al.* G protein-coupled receptors in *Anopheles gambiae. Science* **298,** 176–178 (2002).
17. Dobritsa, A. A. van der Goes van Naters, W. Warr, C. G., Steinbrecht, R. A. & Carlson, J. R. Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. *Neuron* **37,** 827–841 (2003).
18. Benton, R., Sachse, S., Michnick, S. W. & Vosshall, L. B. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors *in vivo. PLoS Biol.* **4,** e20 (2006).
19. Hallem, E. A. & Carlson, J. R. Coding of odors by a receptor repertoire. *Cell* **125,** 143–160 (2006).
20. Shanbhag, S. R., Mueller, B. & Steinbrecht, R. A. Atlas of olfactory organs of *Drosophila melanogaster*. 1. Types, external organization, innervation and distribution of olfactory sensilla. *Int. J. Insect Morphol. Embryol.* **28,** 377–397 (1999).
21. Gong, W. J. & Golic, K. G. Ends-out, or replacement, gene targeting in *Drosophila. Proc. Natl Acad. Sci. USA* **100,** 2556–2561 (2003).
22. Larsson, M. C. *et al. Or83b* encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron* **43,** 703–714 (2004).
23. Gray, J. M. *et al.* Oxygen sensation and social feeding mediated by a *C. elegans* guanylate cyclase homologue. *Nature* **430,** 317–322 (2004).
24. Wingrove, J. A. & O'Farrell, P. H. Nitric oxide contributes to behavioral, cellular, and developmental responses to low oxygen in *Drosophila. Cell* **98,** 105–114 (1999).
25. Verma, A., Hirsch, D. J., Glatt, C. E., Ronnett, G. V. & Snyder, S. H. Carbon monoxide: a putative neural messenger. *Science* **259,** 381–384 (1993).
26. Reinking, J. *et al.* The *Drosophila* nuclear receptor e75 contains heme and is gas responsive. *Cell* **122,** 195–207 (2005).
27. Hou, S. *et al.* Myoglobin-like aerotaxis transducers in Archaea and Bacteria. *Nature* **403,** 540–544 (2000).
28. Laissue, P. P. *et al.* Three-dimensional reconstruction of the antennal lobe in *Drosophila melanogaster. J. Comp. Neurol.* **405,** 543–552 (1999).
29. Fishilevich, E. *et al.* Chemotaxis behavior mediated by single larval olfactory neurons in *Drosophila. Curr. Biol.* **15,** 2086–2096 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Howell and M. Q. Benedict of the CDC and MR4 for providing us with mosquitoes contributed by W. E. Collins; K. Kay and K. Fishilevich for technical assistance; and R. Axel, C. Bargmann, K. J. Lee and members of the Vosshall Laboratory for comments on the manuscript. This work was funded in part by a grant to R. Axel and L.B.V. from the Foundation for the National Institutes of Health through the Grand Challenges in Global Health Initiative and by an NIH grant to L.B.V. Support was contributed to W.D.J. from an NIH MSTP grant, to P.C. from the Jane Coffin Childs Memorial Fund for Medical Research and to I.G.K. from the Human Frontier Science Program.

Author Contributions W.D.J. carried out all the experiments and analysed the data. P.C. and I.G.K. generated and characterized the *Gr63a-sytRFP* transgene in the laboratory of S. L. Zipursky at UCLA. W.D.J. and L.B.V. together designed the experiments, interpreted the results, produced the figures, and wrote the paper.

Author Information Genbank accession numbers for *A. gambiae* genes in this paper are: *GPROR7* (AY843205), *GPRGR22* (DQ989011) and *GPRGR24* (DQ989013). Genbank accession numbers for *D. melanogaster* genes in this paper are: *Gr10a* (DQ989010), *Gr21a* (DQ989014) and *Gr63a* (DQ989012). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.B.V. (leslie@mail.rockefeller.edu).

LETTERS

The cellular machinery of *Ferroplasma acidiphilum* is iron-protein-dominated

Manuel Ferrer[1], Olga V. Golyshina[2], Ana Beloqui[1], Peter N. Golyshin[2,3]* & Kenneth N. Timmis[2,3,4]*

Ferroplasma **is a genus of the Archaea, one of the three branches of the tree of life, and belongs to the order Thermoplasmatales (Euryarchaeota), which contains the most acidophilic microbes yet known.** ***Ferroplasma*** **species live in acid mine drainage, acidic pools and environments containing sulphidic ores such as pyrite and characterized by pH values of 0–2 and high concentrations of ferrous iron and other heavy metals[1–3].** ***F. acidiphilum*** **strain Y^T is a chemoautotroph that grows optimally at pH 1.7 and gains energy by oxidizing ferrous iron and carbon by the fixation of carbon dioxide[1]. Here we show that 86% of 189 investigated cellular proteins of** ***F. acidiphilum*** **are iron-metalloproteins. These include proteins with deduced structural, chaperone and catalytic roles, not described as iron-metalloproteins in any other organism so far investigated. The iron atoms in the proteins seem to organize and stabilize their three-dimensional structures, to act as 'iron rivets'. Analysis of proteins of the phylogenetic neighbour** ***Picrophilus torridus*** **and of the habitat neighbour** ***Acidithiobacillus ferrooxidans*** **revealed far fewer and only typical metalloproteins.** ***F. acidiphilum*** **therefore has a currently unique iron-protein-dominated cellular machinery and biochemical phylogeny.**

We showed recently that an α-glucosidase of *F. acidiphilum* is an iron-containing metalloenzyme, a property not known for any other glycoside hydrolase[4]. Because this enzyme has no significant homology with other proteins, we presumed that it represents a new class of glycoside hydrolase. To gain information on the prevalence of iron-containing proteins in *F. acidiphilum* we prepared a culture of the organism, extracted proteins, subjected them to two-dimensional non-denaturing polyacrylamide gel electrophoresis (PAGE), and treated the gel with the chemiluminescent substrate luminal to stain metalloproteins (Fig. 1). Of 107 proteins detected by this procedure, 87 abundant well-separated spots were analysed; from these, 78 unique proteins were unambiguously identified by quadrupole time-of-flight (Q-TOF) mass spectrometry (Table 1). High-resolution inductively coupled plasma mass spectroscopy (ICP-MS) revealed the presence of iron in all 87 protein spot samples at concentrations ranging from 0.1 to 2.1 ng (Table 1). Other metals—Co, Cu, Ca, Mg, Ni, Zn and Mn—were also identified in ten of these, and S was found in three. In 9 of the 13 proteins containing additional metals, the iron content was either greater than or similar to that of the second metal. Although the 78 different iron-metalloproteins identified included 28 typical metalloproteins, the other 50 (indicated by U for unique in Table 1) are proteins that have until now never or rarely been shown to contain iron.

If most metalloproteins of *F. acidiphilum* contain iron, the crucial question arises: What fraction of the entire proteome is represented by the metalloproteins? We prepared a two-dimensional non-denaturing gel of *F. acidiphilum* cellular proteins and stained it with Coomassie blue: about 1,800 polypeptides were resolved (Supplementary Fig. S1). Some 400 spots were sufficiently discrete to be cored, and 203 polypeptides representing 189 different species were unequivocally identified by both matrix-assisted laser desorption ionization–time-of-flight (MALDI–TOF) tandem mass spectrometry (MS/MS) and Q-TOF peptide mapping, and their metal contents were determined (Supplementary Table S2). Of the 189 proteins analysed, 163 contained iron and 26 did not. Many of the iron-metalloproteins of *F. acidiphilum* are housekeeping proteins that in other organisms are not known to contain iron, and in some cases not even a metal (we distinguish between metalloproteins (proteins containing metals as fixed structural elements) and metal-dependent proteins (proteins transiently requiring exogenous metals for activity)), such as DNA ligase, transposases, endonucleases, integrases–recombinases and DNA repair proteins. Some, for example ribosomal proteins, are structural proteins. Of particular interest are DNA/RNA-binding proteins because, to our knowledge, only a few proteins of this type contain iron clusters (RNA methyltransferases[5], DNA glycosylases[6] and enzymes participating in the methylthiolation of the base moiety of adenosine in some transfer RNAs[7]). Therefore, about 86% of 189 *F. acidiphilum* polypeptides selected essentially at

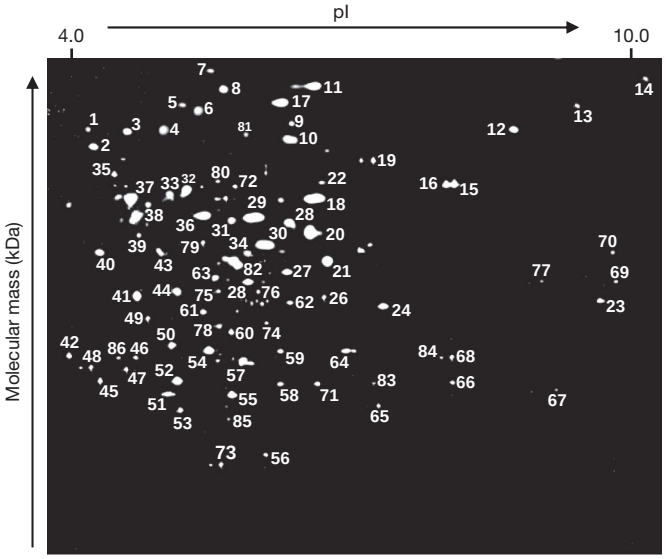


Figure 1 | 'Luminal metalloproteome' of *F. acidiphilum* Y^T. Proteins extracted from exponentially grown cells of *F. acidiphilum* were fractionated by two-dimensional non-denaturing PAGE and subsequently stained by peroxidation of gels with the chemiluminescent substrate luminal to identify metalloproteins.

[1]CSIC, Institute of Catalysis, Cantoblanco, 28049 Madrid, Spain. [2]Department of Environmental Microbiology, HZI-Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany. [3]Institute of Microbiology, Carolo-Wilhelmina Technical University of Braunschweig, 38106 Braunschweig, Germany. [4]Department of Biological Sciences, University of Essex, Colchester CO4 3SQ, UK.
*These authors contributed equally to this work.

Table 1 | Metalloproteins identified by ESI-Q-TOF sequencing in the 'luminal metalloproteome' of *F. acidiphilum* Y^T

Spot no.*	Molecular mass (kDa) (monomer)†	pI†	Protein identification‡	Metal identification (quantity, ng)§	Comment‖
1	58.00	5.19	Thermosome β subunit	Fe (0.57)	U
2	54.25	9.79	Transposase	Fe (2.10)	U
3	67.90	5.31	LigFa, ATP-dependent DNA ligase (EC 6.5.1.1)	Fe (0.60)	U
4	68.22	5.57	Pyruvate decarboxylase	Fe (0.32), Cu (1.20)	C
5	66.20	6.74	4Fe–4S ferredoxin (EC 1.2.7.8)	Fe (0.42), S (0.21)	C
6	75.19	5.79	Glutamate synthase (NADPH) α subunit (EC 1.4.1.13)	Fe (0.68)	C
7	183.95	5.90	Putative ATP-dependent helicase	Fe (0.47)	U
8	72.32	5.64	Acetoacetyl-CoA synthetase (EC 6.2.1.16)	Fe (0.33), Co (0.54)	U
9	63.18	5.51	Acyl-CoA synthase (EC 6.2.1.3)	Fe (0.15), Mg (0.24)	U
10	35.25	6.30	Formate hydrogenlyase (EC 1.2.1.2)	Fe (0.53), Ni (0.24)	C
11	96.60	6.11	Calcium transporter (ATPase) (EC 3.6.3.8)	Fe (0.32), Mg (0.10)	U
12	72.02	7.62	2-Oxoacid ferrodoxin oxidoreductase (αsubunit)	Fe (0.30)	C
13	85.16	8.95	DNA topoisomerase I (EC 5.99.1.2)	Fe (0.58), Ca (0.26)	U
14	70.29	9.15	Cytochrome *c* oxidase subunit I (EC 1.9.3.1)	Fe (0.41), Zn (0.15)	C
15	54.66	7.06	4-Hydroxyphenylacetate 3-monooxygenase (EC 1.14.13.3)	Fe (0.36)	U
16=15	54.66	7.06	4-Hydroxyphenylacetate 3-monooxygenase (EC 1.14.13.3)	Fe (0.26)	U
17	73.58	6.54	Copper-translocating P-type ATPase (EC 3.6.3.4)	Fe (0.29)	C
18	58.27	5.56	Thermosome α subunit	Fe (0.76)	U
19	66.23	6.74	RNase L inhibitor homologue, predicted ATPase	Fe (0.28), S (0.12)	C
20	49.50	6.61	D-Hydantoinase (dihydropyrimidinase) (EC 3.5.2.2)	Fe (0.90)	C
21	40.51	6.42	Muconate cycloisomerase-related protein (EC 5.5.1.1)	Fe (0.30)	C
22	62.72	6.65	Acyl-CoA ligase (AMP-binding) (EC 6.2.1.3)	Fe (0.20)	U
23	43.87	9.21	Geranylgeranyl reductase	Fe (0.16)	U
24	44.80	5.61	Hydroxymethylglutaryl-CoA reductase (EC 1.1.1.34)	Fe (0.87)	U
25	47.72	6.73	Adenosylhomocysteinase (EC 3.3.1.1)	Fe (0.24)	U
26=25	47.72	6.73	Adenosylhomocysteinase (EC 3.3.1.1)	Fe (0.26)	U
27	44.62	6.28	Asparaginase (EC 3.5.1.1)	Fe (1.15)	U
28	50.66	6.30	Benzoylformate decarboxylase (EC 4.1.1.7)	Fe (1.28)	U
29	53.27	6.20	Glycine dehydrogenase (subunit 2) (EC 1.4.4.2)	Fe (0.54)	U
30	47.65	6.21	NADH oxidase (cytochrome *c* reductase) (EC 1.6.99.3)	Fe (1.40), S (6.1)	C
31	51.92	5.84	Glutamate decarboxylase isoenzyme 1 (EC 4.1.1.15)	Fe (0.71)	U
32=9	63.18	5.51	Acetyl-CoA synthase (EC 6.2.1.3)	Fe (0.54), Mg (0.83)	U
33	61.23	5.44	Fatty-acid CoA ligase (EC 6.2.1.3)	Fe (0.15)	U
34	46.02	6.22	Phosphoglycerate kinase (EC 2.7.2.3)	Fe (0.91)	U
35	61.01	5.43	Pyruvate oxidase (EC 1.2.2.2)	Fe (0.34)	C
36	50.52	5.58	Mercuric reductase (EC 1.16.1.1)	Fe (0.64)	U
37	54.27	5.52	NAD^+-dependent aldehyde dehydrogenase (EC 1.2.1.3)	Fe (0.55)	C
38	53.28	5.40	Glutamyltransferase (EC 2.3.2.2)	Fe (1.47)	U
39	46.89	5.49	Argininosuccinate lyase (EC 6.3.4.5)	Fe (0.13)	U
40	50.45	5.14	Glutamine synthethase (EC 6.3.1.2)	Fe (0.35)	U
41	44.61	5.46	Formate hydrogenlyase (subunit III) (EC 1.2.1.2)	Fe (1.70), Ni (1.35)	C
42	48.06	6.40	Glycine/serine hydroxymethyltransferase (EC2.1.2.1)	Fe (0.51)	U
43	50.24	5.52	Argininosuccinate synthase (EC 6.3.4.5) (not the same as spot 39)	Fe (0.30)	U
44	46.05	5.50	Putative galactonate dehydratase (EC 4.2.1.6)	Fe (1.20)	U
45	32.45	5.60	Methionine aminopeptidase (EC 3.4.11.18)	Fe (0.36)	C
46	37.94	4.71	Phosphoribosylformylglycinamidine cyclo-ligase (EC 6.3.3.1)	Fe (1.16), Ni (1.23)	U
47	31.64	5.05	Pyruvate ferredoxin oxidoreductase (beta subunit) (EC1.2.7.1)	Fe (0.12)	C
48	43.71	5.23	Cytochrome P450 (EC 1.14.14.1)	Fe (0.14)	C
49	41.88	5.30	Sulphide dehydrogenase–flavocytochrome *c* (EC 1.8.2.-)	Fe (0.93)	C
50	39.02	5.86	Carbamoyl phosphate synthethase small subunit (EC 6.3.5.5)	Fe (1.04)	U
51=45	32.45	5.60	Methionine aminopeptidase (EC 3.4.11.18)	Fe (0.18)	C
52=49	41.88	5.30	Sulphide dehydrogenaseflavocytochrome C (EC 1.8.2.-)	Fe (1.32)	C
53	27.65	5.35	Enoyl-CoA hydratase (EC 4.2.1.17)	Fe (0.56)	U
54	34.59	5.65	EstFa, carboxyl-esterase (EC 3.1.1.1)	Fe (0.39)	U
55	34.92	6.17	Thioredoxin reductase (EC 1.6.4.5)	Fe (0.67)	C
56	23.04	5.85	Cytochrome *c* oxidase subunit II (EC 1.9.3.1)	Fe (0.48)	C
57=10	35.25	6.30	Formate hydrogenlyase (subunit IV) (EC 1.2.1.2)	Fe (0.53), Ni (0.64)	C
58	35.14	6.33	NDP-sugar epimerase (EC 5.1.3.2)	Fe (0.60)	U
59	40.50	6.42	GlyFa2 α-glucosidase (EC 3.2.1.20)	Fe (0.15)	U
60	42.56	6.09	3-Ketoacyl-CoA thiolase (EC 2.3.1.16)	Fe (0.68)	U
61	44.61	5.68	*N*-Carbamyl-L-amino acid amidohydrolase (EC 3.5.1.77)	Fe (0.44)	U
62=27	44.62	6.28	Asparaginase (EC 3.5.1.1)	Fe (0.53)	U
63	20.70	5.10	Protease (ThiJ/PfpI family) (EC 3.4.-.-)	Fe (0.37)	C
64=59	40.50	6.42	GlyFa2, α-glucosidase (EC 3.2.1.20)	Fe (0.51)	U
65	28.62	6.34	2-Dehydro-3-deoxyphosphoheptonate aldolase (EC 2.5.1.54)	Fe (0.28)	C
66	34.23	7.12	Cysteine synthase (EC 2.5.1.47)	Fe (0.46)	U
67	28.32	9.53	GlyFa1, α-glucosidase (EC 3.2.1.20)	Fe (0.31)	U
68	38.49	6.77	*N*-Acetyl-γ-glutamyl-phosphate reductase (EC 1.2.1.38)	Fe (0.23)	C
69	47.03	7.74	Elongation factor TU 1-α subunit (EC 3.6.5.3)	Fe (0.39)	U
70	54.40	9.21	Putative permease	Fe (0.22)	U
71	28.02	6.33	Acetylglutamate kinase (EC 2.7.2.8)	Fe (0.63)	U
72	59.35	6.31	Dihydroxy-acid dehydratase (EC 4.2.1.9)	Fe (0.24)	C
73	24.46	5.99	Fe-superoxide dismutase (EC 1.15.1.1)	Fe (1.21), Mn (0.61)	C
74	39.88	6.11	Cystathionine γ-synthase (EC 4.4.1.8)	Fe (1.03)	U
75	45.50	5.74	Glutamate dehydrogenase (EC 1.4.1.3)	Fe (0.25)	U
76	45.40	6.16	3-Isopropylmalate dehydratase large subunit (EC 4.2.1.33)	Fe (0.57)	C
77	47.69	7.61	Malate oxidoreductase (EC 1.1.1.38)	Fe (0.95)	U
78	40.99	5.95	ATP synthase subunit C (EC 3.6.3.15)	Fe (0.71)	U
79	48.90	5.56	4-Aminobutyrate transaminase (EC 2.6.1.19)	Fe (0.27)	U
80=72	59.35	6.31	Dihydroxy-acid dehydratase (EC 4.2.1.9)	Fe (0.86)	C
81	67.26	6.17	Peptidase S9, prolyl oligopeptidase (EC 3.4.21.26)	Fe (0.30)	U
82	47.65	6.21	NADH dehydrogenase (EC 1.6.99.3)	Fe (0.70)	C
83	29.85	6.93	Decaprenyl-diphosphate synthase (EC 2.5.1.31)	Fe (1.21)	U
84	37.62	7.08	Threonine synthase (EC 4.2.99.2)	Fe (0.15)	U
85	23.46	6.12	Triosephosphate isomerase (EC 5.3.1.1)	Fe (0.08)	U
86	41.32	5.27	4-Hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27)	Fe (0.23)	C
87	57.30	6.47	αGluFa, α-glucosidase (EC 3.2.1.20)	Fe (0.73)	U

* An equals sign indicates that proteins corresponding to these spots are identical.
† Monomeric molecular mass and pI of all proteins were calculated from the theoretical molecular mass of the most closely homologous protein from '*F. acidarmanus*' found after BlastP searches for 'short, nearly exact matches'.
‡ *Ferroplasma* proteins are metal-containing proteins. Identification was made by ESI-Q-TOF sequencing, coupled with online database searching for homologous proteins, using BLASTX and BLASTP tools with default settings at NCBI (http://www.ncbi.nlm.nih.gov/blast), and BlastP searches for 'short, nearly exact matches' (for details see Supplementary Table S1).
§ For metal content determination the spots were excised, digested with 0.5% (v/v) nitric acid and analysed by ICP-MS. The results shown are the averages of data derived from three different experiments ($n = 3$). The three sets of data are statistically equivalent and standard deviations are not shown.
‖ U, unique (homologous proteins have never been reported to contain iron); C, common (iron is a common cofactor in homologous proteins or is present in at least one homologous protein).

random were iron-metalloproteins. By extrapolation, about 86% of the entire protein repertoire of *Ferroplasma* may consist of iron-metalloproteins. *F. acidiphilum* is therefore a cellular entity characterized by a currently unique iron-metalloprotein-dominated metabolic machinery.

To determine whether iron is present in stoichiometric amounts in *F. acidiphilum* proteins, we quantified proteins and metals in individual polypeptide spots from five stained gels. As shown in Supplementary Table S2, the protein and iron amounts varied from about 22.0 ± 0.9 to 638.4 ± 28.1 ng and from 110 ± 9 to $9{,}300 \pm 360$ pg, respectively, giving ratios of 0.2–33 mol of iron per monomer, with all except 20 proteins exhibiting stoichiometries greater than 1. These stoichiometries are consistent with the notion that iron is bound to *F. acidiphilum* proteins in a specific manner, coordinated by iron ligands in defined domains, and is functional. To further assess whether iron is essential in *Ferroplasma* proteins, we removed it from six purified proteins, selected essentially at random (accurately determined iron:protein stoichiometries were all well correlated with those estimated by spot analyses), and measured their activities and circular dichroism spectra (Table 2). Because we do not know whether iron in these proteins is in the Fe^{2+} or Fe^{3+} form, and Fe^{3+} is difficult to remove, we first converted the iron in the proteins to Fe^{2+} by reduction with dithionite, and then removed the metal with a chelating agent. Decreases in iron content of more than 90% were achieved in all cases, and these decreases were accompanied by more than 80% decreases in ellipticity, according to circular dichroism measurements, and thus protein secondary structure, and 80% or greater decreases in activity. Iron is therefore crucial for maintenance of the three-dimensional structure, and hence the activity, of the proteins examined, a function we designate as the 'iron rivet'.

These findings immediately provoke the question: Do other microbes, either phylogenetically related to *F. acidiphilum* or inhabiting the same environment, also have such an iron-dominated metalloprotein repertoire? We therefore analysed the metalloproteomes of the acidophilic and closely related archaeon *Picrophilus torridus* and of the unrelated *Acidithiobacillus ferrooxidans*, which belongs to the Bacteria, the second branch of the tree of life, and which occupies ecological niches similar to those of *F. acidiphilum*[2]. By means of staining with luminal and mapping by Q-TOF MS, we consistently identified in the metalloproteomes of *P. torridus* and *A. ferrooxidans* 29 and 35 spots, representing 29 and 28 unique proteins, respectively (Supplementary Fig. S2). About one-half of these proteins contained Fe, about one-third contained Zn, about one-third contained S (mostly in combination with Fe), and the others contained Mg, Mo, Ca, Hg, Cu or Ni. Eleven metalloproteins of *P. torridus* and eight of *A. ferrooxidans* contained both Fe and another metal. Most significantly, all Fe-containing proteins detected by this procedure were typical iron-metalloproteins found in other organisms.

What could be the explanation for the iron-metalloprotein-dominated metabolic machinery of *F. acidiphilum?* Iron is the fourth most abundant element in the Earth's crust and is crucial to diverse catalytic, metabolic and physiological functions. However, it is weakly soluble and poorly bioavailable to cellular systems (it is often the factor limiting microbial growth/productivity in the open ocean[8,9]); in consequence, cells invest considerable metabolic resources in acquiring it[10]. This must have constituted a powerful selective force for the evolution of diverse iron-independent cellular mechanisms that is reflected in the current predominance of metal-free proteins, and in the evolution of metalloproteins containing metals other than iron. An exception to the global iron-limited biosphere is acidic iron-rich environments, inhabited by *Ferroplasma*, in which iron is highly soluble and not limiting for cellular needs. However, the simple conclusion that iron-based protein functions that characterize *Ferroplasma* are superior to others, and that organisms living in iron-rich habitats are unconstrained in exploiting iron for such functions, is not supported by the absence of similar iron-protein repertoires in *Acidithiobacillus*, which inhabits similar environments. Thus, neither habitat ecophysiological nor phylogenetic considerations provide any useful clues about the 'iron rivet' of *F. acidiphilum*.

One intriguing possibility we have considered is that the 'iron rivet' is an ancient property that has uniquely (in terms of currently known cellular systems) been retained by *F. acidiphilum*. In this regard, it is interesting to note that one current theory of the origin of life invokes iron–sulphur chemistry catalysing the formation and further transformation of the organic molecules of life, and taking place on iron–sulphur-rich surfaces such as pyrite[11,12]. Iron–sulphur-mediated catalysis evolved into protein catalysis, some of which retained iron–sulphur chemistry. The transition to flexible high-molecular-mass catalysts, with complex three-dimensional structures and folds, would have required the evolution of structure-organizing and structure-stabilizing elements. Perhaps, therefore, multivalent iron that was in abundance in the sites where life may have evolved became an early element of the organization and stabilization of protein structure ('iron rivets' holding together complex, inherently fragile structures), a forerunner of other structure-stabilizing elements. In this model, the escape of early cellular forms from the pyrite environment to diverse environments characterized by poor iron bioavailability would have been a powerful selective force for the evolution of iron-independent proteins. As these evolved, natural selection would have eliminated the less fit mechanisms and iron would have been retained only for functions in which it cannot be effectively replaced by alternative mechanisms (presumably iron–sulphur redox centres and haem, for example)[13,14]. Acidic pyrite environments on the Earth's surface still exist as iron–sulphur-rich habitats[2,3], extreme environments inhabited by low-diversity microbial communities and lacking any obvious selective pressure for the evolution of iron-independent alternative catalytic systems. The possibility therefore arises that, unlike other habitat and phylogenetic neighbours, the *Ferroplasma* lineage might have evolved entirely within the pyrite habitat and its protein repertoire might represent an extant relic of early cellular life. This would predict that

Table 2 | Effect of removal of iron from *F. acidiphilum* proteins

Parameter	Iron*	Enzyme					
		α-GluFa	GlyFa1	GlyFa2	EstFa	LigFa	Thermosome
Iron content (mol Fe per mol protein)*†	+	1.04	4.08	2.02	1.03	2.03	32.40
	−	0.05	0.17	0.16	0.02	0.07	1.10
$[\theta]_{220\,nm} \times 10^{-4}$ (deg cm^2 dmol^{-1})*	+	−4.50	−4.70	−5.70	−7.30	−7.10	−6.60
	−	−0.09	−0.34	−0.93	−1.10	−0.80	−1.25
Activity*‡	+	293.0	142.0	142.0	64.2	96,000	230
	−	5.9	18.5	22.7	2.6	1,920	46

* The enzymes analysed were three α-glucosidases (α-GluFa, GlyFa1 and GlyFa2), one esterase (EstFa), one DNA ligase (LigFa) and the thermosome complex (archaeal chaperonins containing 8 α and 8 β subunits). For protein purification and assays of iron-containing (+) versus iron-depleted (−) enzymes see Supplementary Methods. All experiments were performed in triplicate (average values are shown, with standard deviation being less than 5%).

† As shown, the experimental iron:protein stoichiometries correlated perfectly well with those obtained by spot analysis (see Supplementary Table S2), indicating that the stoichiometries obtained by spot analyses are reasonably accurate.

‡ For α-glucosidases and esterase the activity is referred to the kinetic parameter k_{cat}/K_m (in s^{-1} mM^{-1}) towards sucrose and *p*-nitrophenyl propionate, respectively. DNA ligase was assayed with *Sau3a*-digested λ DNA and is given as units mg^{-1} (one unit being defined as the amount of enzyme required for ligation of 50% of DNA substrate in 30 min and corresponds to 0.015 Weiss units). For thermosome, the activity is given as μmol of ATP hydrolysed per complex per minute. Activity assays were performed at 40 °C in 100 mM sodium citrate buffer, pH 3.0.

at least some of the few iron-lacking proteins of *F. acidiphilum* will have evolved from genes acquired by horizontal transfer from habitat partners that evolved from microbes on a non-pyrite evolutionary trajectory (for example *A. ferrooxidans*) and that the majority iron-containing protein repertoire uniquely features a primitive structure organizing and stabilizing mechanism, the 'iron rivet'. If this turns out to be true, *Ferroplasma* will open a new window into early life and its evolutionary development.

METHODS

Strains *F. acidiphilium* Y^T (DSM 12658^T), *P. torridus* (DSM9790^T) and *A. ferrooxidans* (ATCC23270; DSM14882^T) were obtained from the DSMZ strain collection and harvested in the exponential/late-exponential growth phase: *F. acidiphilum* Y^T was cultured at 37 °C for 96 h in 9K medium supplemented with 0.02% yeast extract, pH 1.7, as described previously[1]; *A. ferrooxidans* was grown at 30 °C for 72 h in 9K medium, pH 2.0; and *P. torridus* was cultured at 55 °C for 72 h in medium 88 (http://www.dsmz.de/microorganisms/html/media/medium000088.html) supplemented with 1 g l^{-1} glucose, 0.1% yeast extract and, where indicated, 5 g l^{-1} of $FeSO_4.7H_2O$. Two-dimensional gel electrophoresis under non-denaturing conditions was performed with 300 μg of proteins, isolated from frozen cells as described elsewhere[15], on 24-cm pH 3–10 NL IPG strips (ReadyStrip; Bio-Rad) and 1.5-mm-thick 10–15% non-denaturing polyacrylamide gels as described elsewhere[16]. Detection of (metallo)proteins was performed either by direct staining with the chemiluminescent substrate luminal[17] or by a more comprehensive approach involving normal Coomassie blue staining of gels. Protein spots were cored from preparative gels and identified by trypsin digestion *in situ* and MALDI-TOF MS/MS (REFLEX; Brucker) and hybrid Q-TOF MS with MS/MS capability (ESI-Q-TOF MS/MS; QSTAR; Applied Biosystems) coupled with HPLC 1100 (Agilent Technologies). The metal ion content of cored protein spots was measured by ICP-MS, after dilution with 0.5 ml of 0.5% (v/v) HNO_3 to digest the protein and release the metal ions. All proteins from *F. acidiphilum* used in the present study were produced and purified as iron-containing and iron-free forms and analysed as described previously[4,16]. Full details of all methods are given in Supplementary Methods.

Received 26 June; accepted 20 October 2006.

1. Golyshina, O. V. *et al. Ferroplasma acidiphilum* gen. nov., sp. nov., an acidophilic, autotrophic, ferrous-iron-oxidizing, cell-wall-lacking, mesophilic member of the *Ferroplasmaceae* fam. nov., comprising a distinct lineage of the *Archaea*. *Int. J. Syst. Evol. Microbiol.* **3,** 997–1006 (2000).
2. Golyshina, O. V. & Timmis, K. N. *Ferroplasma* and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments. *Environ. Microbiol.* **7,** 1277–1288 (2005).
3. Dopson, M., Baker-Austin, C., Hind, A., Bowman, J. P. & Bond, P. C. Characterization of *Ferroplasma* isolates and *Ferroplasma acidarmanus* sp. nov., extreme acidophiles from acid mine drainage and industrial bioleaching environments. *Appl. Environ. Microbiol.* **70,** 2079–2088 (2004).
4. Ferrer, M., Golyshina, O. V., Plou, F. J., Timmis, K. N. & Golyshin, P. N. A novel α-glucosidase from the acidophilic archaeon *Ferroplasma acidiphilum* strain Y with high transglycosylation activity and an unusual catalytic nucleophile. *Biochem. J.* **391,** 269–276 (2005).
5. Agarwalla, S., Stroud, R. M. & Gaffney, B. J. Redox reactions of the iron–sulfur clusters in a ribosomal RNA methyltransferase, RumA: optical and EPR studies. *J. Biol. Chem.* **279,** 34123–34129 (2004).
6. Boal, A. K. *et al.* DNA-bound redox activity of DNA repair glycosylases containing [4Fe-4S] clusters. *Biochemistry* **44,** 8397–8407 (2005).
7. Pierrel, F., Bjork, G. R., Fontecave, M. & Atta, M. Enzymatic modification of tRNAs: MiaB is an iron–sulfur protein. *J. Biol. Chem.* **277,** 13367–13370 (2002).
8. Watson, A. J., Bakker, D. C., Ridgwell, A. J., Bord, P. W. & Law, C. S. Effect of iron supply on Southern Ocean CO_2 uptake and implications for glacial atmospheric CO_2. *Nature* **407,** 730–733 (2000).
9. Buesseler, K. O., Andrews, J. E., Pike, S. M. & Charetter, M. A. The effect of iron fertilization on carbon sequestration in the Southern Ocean. *Science* **304,** 414–417 (2004).
10. Imlay, J. A. Iron–sulphur clusters and the problem with oxygen. *Mol. Microbiol.* **59,** 1073–1082 (2006).
11. Blochl, E., Keller, M., Wächterhauser, G. & Stetter, K. O. Reactions depending on iron sulfde and linking geochemistry with biochemistry. *Proc. Natl Acad. Sci. USA* **89,** 8117–8120 (1992).
12. Wächtershäuser, G. Discussing the origin of life. *Science* **296,** 1982–1983 (2002).
13. Major, T. A., Burd, H. & Whitman, W. B. Abundance of 4Fe-4S motifs in the genomes of methanogens and other prokaryotes. *FEMS Microbiol. Lett.* **239,** 117–123 (2004).
14. Schneider, D. & Schmidt, C. L. Multiple Rieske proteins in prokaryotes: where and why? *Biochim. Biophys. Acta* **1710,** 1–12 (2005).
15. Giometti, C. S. *et al.* Analysis of the *Shewanella oneidensis* proteome by two-dimensional gel electrophoresis under non-denaturing conditions. *Proteomics* **3,** 777–785 (2003).
16. Golyshina, O. V., Golyshin, P. N., Timmis, K. N. & Ferrer, M. The 'pH optimum anomaly' of intracellular enzymes of *Ferroplasma acidiphilum*. *Environ. Microbiol.* **8,** 416–425 (2006).
17. Högbom, M. *et al.* A high-throughput method for the detection of metalloproteins on a milligram scale. *Mol. Cell. Proteomics* **4,** 827–834 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Wächtershäuser, R. Thauer, M. Wilson, A. Böck and A. Ballesteros for discussions. This research was supported by the Spanish Ministerio de Educación y Ciencia (MEC) (Ramón y Cajal contract to M.F. and a FPU Fellowship to A.B.), European Community Project 'BIOMELI', the Microbial Genomic Network Programme (GenoMik Plus) of the German Ministry for Education and Research (BMBF), and the DFG Priority Program 'Mars and Terrestrial Planets'. K.N.T. thanks the Fonds der Chemischen Industrie for generous support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.F. (mferrer@icp.csic.es).

LETTERS

A systems biology analysis of the *Drosophila* phagosome

L. M. Stuart[1,2], J. Boulais[3], G. M. Charriere[1], E. J. Hennessy[1], S. Brunet[3], I. Jutras[3], G. Goyette[3], C. Rondeau[3], S. Letarte[3], H. Huang[4], P. Ye[4], F. Morales[5], C. Kocks[1], J. S. Bader[4], M. Desjardins[3] & R. A. B. Ezekowitz[1]†

Phagocytes have a critical function in remodelling tissues during embryogenesis and thereafter are central effectors of immune defence[1,2]. During phagocytosis, particles are internalized into 'phagosomes', organelles from which immune processes such as microbial destruction and antigen presentation are initiated[3]. Certain pathogens have evolved mechanisms to evade the immune system and persist undetected within phagocytes, and it is therefore evident that a detailed knowledge of this process is essential to an understanding of many aspects of innate and adaptive immunity. However, despite the crucial role of phagosomes in immunity, their components and organization are not fully defined. Here we present a systems biology analysis of phagosomes isolated from cells derived from the genetically tractable model organism *Drosophila melanogaster* and address the complex dynamic interactions between proteins within this organelle and their involvement in particle engulfment. Proteomic analysis identified 617 proteins potentially associated with *Drosophila* phagosomes; these were organized by protein–protein interactions to generate the 'phagosome interactome', a detailed protein–protein interaction network of this subcellular compartment. These networks predicted both the architecture of the phagosome and putative biomodules. The contribution of each protein and complex to bacterial internalization was tested by RNA-mediated interference and identified known components of the phagocytic machinery. In addition, the prediction and validation of regulators of phagocytosis such as the 'exocyst'[4], a macromolecular complex required for exocytosis but not previously implicated in phagocytosis, validates this strategy. In generating this 'systems-based model', we show the power of applying this approach to the study of complex cellular processes and organelles and expect that this detailed model of the phagosome will provide a new framework for studying host–pathogen interactions and innate immunity.

Proteomic analysis of the mammalian phagosome has highlighted the complexity of this subcellular organelle[5–7]. To probe this complexity further we chose to study *Drosophila* S2 cells, an embryonic haemocyte-derived cell line that is readily amenable to RNA-mediated interference (RNAi) and whose phagocytic properties have been extensively characterized[8,9] (see Supplementary Fig. S1 for a diagram of the approach). We first undertook a proteomic analysis of latex-bead-containing phagosomes isolated from S2 cells that were highly enriched for phagosome components (estimated contamination less than 5%; see Supplementary Methods and Supplementary Figs S2–S4). SDS polyacrylamide gelelectrophaesis and tandem mass spectrometry analysis identified 617 proteins potentially associated with *Drosophila* phagosomes (Supplementary Table S1 and Supplementary Methods). Of these phagosome proteins, 122 (19.8%) were predicted to have transmembrane domains and 103 were previously undefined (defined only by the prefix CG, for computed gene). The phagosome components were classified, identifying orthologues in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, mouse and humans. Mammalian orthologues were identified for 70% of the *Drosophila* phagosome proteins (Supplementary Table S1). Of the 140 proteins previously identified in the mammalian phagosome, 100 (70%) had orthologues within the *Drosophila* phagosome[5–7], indicating that S2 cells were a valid system from which to derive a model of phagocytosis.

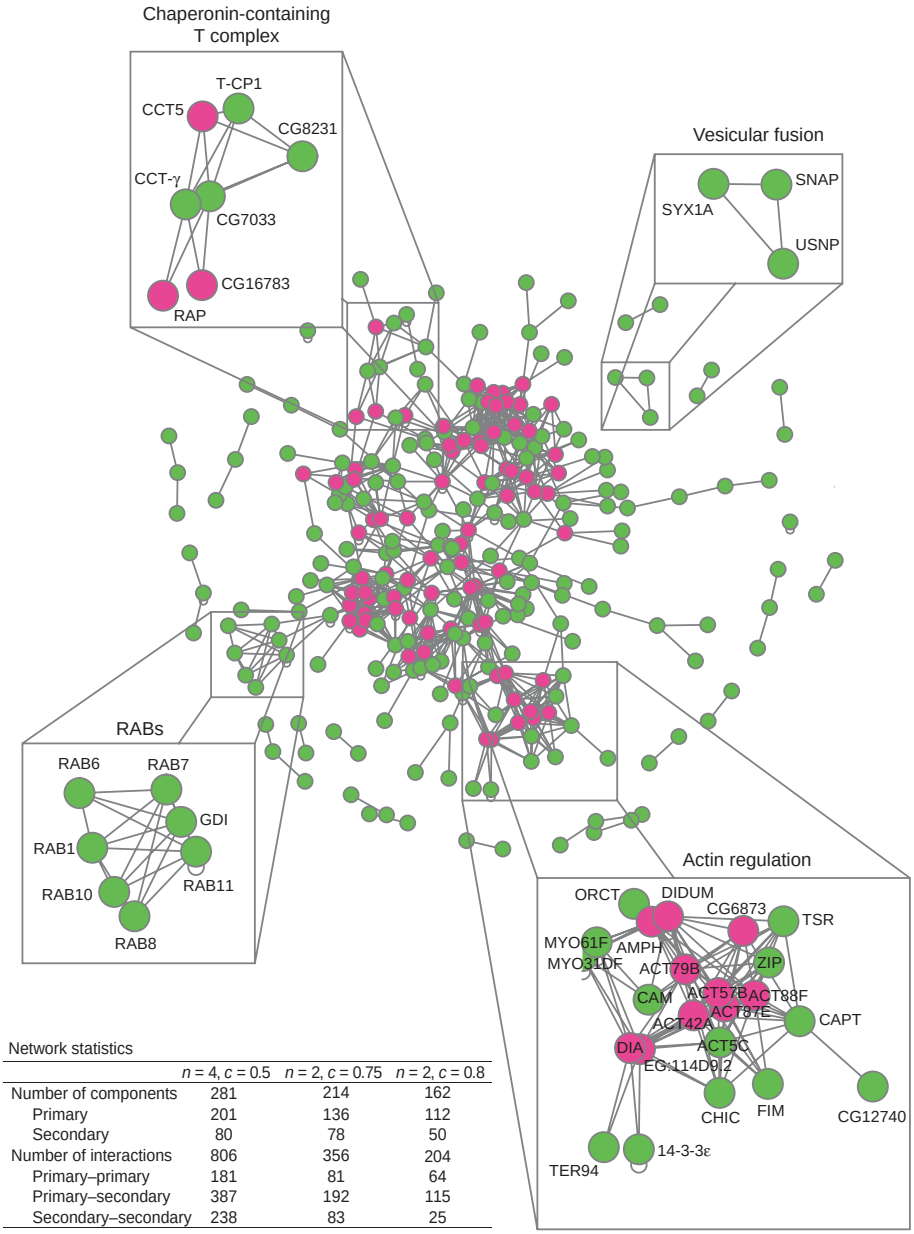
Proteomic analyses of phagosomes have been enlightening but have limitations when interpreted in isolation[10]. First, the false negative rates are unknown but may be significant. Pertinent to this organelle, the data are derived from a single time point in the dynamic evolution of maturing phagosomes and hence might fail to include proteins transiently recruited either as the phagosome forms or while it undergoes maturation. Second, because of the sensitivity of mass spectrometry methods, minor contaminants that purify together with phagosomes may be identified as false positives. Third, proteomics provides no information about the physical or functional organization of the phagosome components. We therefore chose to apply a systems-based analysis to evaluate and organize these proteomic data further (Supplementary Fig. S1).

To place the phagosome proteins in biological context, the proteomic data were used to generate a protein–protein interaction network of this organelle[11–13]. A base network of protein–protein interactions was first generated from confidence-rated interactions observed in five large-scale, high-throughput, experimental screens (Supplementary Methods). Mapping interactions across species provided an additional advantage because it predicted 'interologues' (protein interactions between orthologous proteins in different species) and facilitated aligning protein interaction networks[14,15]. Statistical confidence in a predicted interaction was calculated from the confidence of observed orthologous interactions across species (Supplementary Methods)[13]. This approach yielded a base network incorporating 53,775 observed and inferred protein–protein interactions and associations of complexes in *Drosophila*. The 617 primary component proteins identified in the phagosome proteomics were then used to anchor a network of protein–protein interactions within this larger network. Various methods have been reported for exploratory analysis of confidence-weighted or probabilistic protein interaction networks, with a theme of extending a network through high-probability links[12,16]. We adopted a modified approach on the basis of the expectation that the primary components identified by

[1]Laboratory of Developmental Immunology, Massachusetts General Hospital/ Harvard Medical School, 55 Fruit Street, Boston, Massachusetts 02114, USA. [2]MRC Centre for Inflammation Research, The University of Edinburgh, The Queen's Medical Research Institute, 47 Little France Crescent, Edinburgh EH16 4TJ, UK. [3]Département de pathologie et biologie cellulaire, Université de Montréal, Montréal, Québec H3C 3J7, Canada. [4]Department of Biomedical Engineering and High-Throughput Biology Center, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA. [5]Department of Biomedical Engineering, Faculty of Medicine, McGill University, Montreal, Quebec H3A 2B4, Canada. †Present address: Merck Research Laboratories, RY80K-107, PO Box 2000, Rahway, New Jersey 07065, USA.

experimental proteomics should remain central to the network. Using this strategy we were able to map 214 of the 617 phagosome proteins with high confidence. This map was then extended by prediction of secondary components, proteins absent from the proteomics but predicted from our interaction mapping as potentially associated with the phagosome (Supplementary Methods). The densest network contained 281 total vertices (primary and secondary components) and 806 edges (interactions) (Fig. 1). In addition, the highly sensitive network is shown in Fig. 2 in which clear clusters of potential functional complexes can be defined.

The building of protein–protein interaction networks complemented our original proteomic data in two ways. First, by extending the 'starting players' (primary components) we identified 50–80 secondary components (Supplementary Table S7). These include an unconventional myosin (*didum*), *Amph* (an amphiphysin orthologue) and *diaphanous* (Figs 1 and 2), all implicated in phagocytosis in certain mammalian systems[17–19]. These represent potential false negatives not identified by tandem mass spectrometry perhaps because they are recruited to the phagosome transiently or only with specific cargo and are therefore absent from the latex-bead phagosomes used for proteomics. Second, we reasoned that these networks would provide information concerning the organization of the phagosome. The networks were in keeping both with a global interconnectivity throughout the organelle and with the organization of the phagosome into protein complexes and functional biomodules (identified as topologically interconnected proteins) (Figs 2 and



Network statistics

	$n = 4, c = 0.5$	$n = 2, c = 0.75$	$n = 2, c = 0.8$
Number of components	281	214	162
Primary	201	136	112
Secondary	80	78	50
Number of interactions	806	356	204
Primary–primary	181	81	64
Primary–secondary	387	192	115
Secondary–secondary	238	83	25

Figure 1 | Generation of the 'phagosome interactome'. The network of high-confidence predicted protein–protein interactions between primary phagosome components was generated from yeast two-hybrid screens from *Drosophila* and orthologous proteins in *C. elegans* and yeast as well as using the hybrigenics yeast mass spectrometry pull-down data set. The network corresponding to $(n,c) = (4,0.5)$ is shown. Only high-confidence interactions ($c > 0.5$) are shown, and vertices with no predicted interactions are omitted for clarity. Green vertices represent 'primary components' (that is, those proteins identified by proteomics). Pink vertices are 'secondary components' (proteins predicted to interact by at least four high-confidence connections to primary phagosome proteins (that is, $n = 4$) but not present in our proteomic analysis). Higher-resolution maps with gene names are available on the Developmental Immunology website (http://www.massgeneral.org/devimmunol/phagosome/). Network statistics for three values of (n,c) are tabulated.

3d). Many of these were also confirmed with Gene Ontology (GO) as an independent means of classification (see Supplementary Tables S2–S5 and Supplementary Fig. S5).

Finally, to ascribe function to the phagosome components and to begin to validate the interaction map, we used RNAi and an assay based on fluorescence-activated cell sorting[9,20] to screen 837 genes (including the 617 identified by proteomics) for their role in the phagocytosis of the Gram-positive and Gram-negative pathogens *Staphylococcus aureus* and *Escherichia coli* (Supplementary Table S6). In addition, we tested the role of the secondary components (Supplementary Table S7); 28% of the RNAi treatments affected phagocytosis, either increasing or decreasing bacterial uptake. This represented a fivefold enrichment over our previous screen, in which randomly generated double-stranded RNAs were used[20] ($P = 2 \times 10^{-40}$ for enrichment; see Supplementary Table S9 for hit-rate enrichment analyses). Using this RNAi strategy we identified genes previously implicated in phagocytosis, including those encoding Rac, Cdc42, Rabs and proteins regulating vesicle trafficking. Although some were required for uptake of both organisms, we also identified components involved in the internalization of specific organisms (for example, the engulfment of *S. aureus*, but not that of *E. coli*, occurred primarily by a Rac2/RhoL-dependent mechanism). Genes were classified into three groups by using a more than 30% decrease as our hit limit: first, 31 genes decreased the internalization of both *S. aureus* and *E. coli* (23 with *z* scores of less than −1.5); second, 34 genes decreased only *S. aureus* internalization (44 with *z* scores of less than −1.5); and

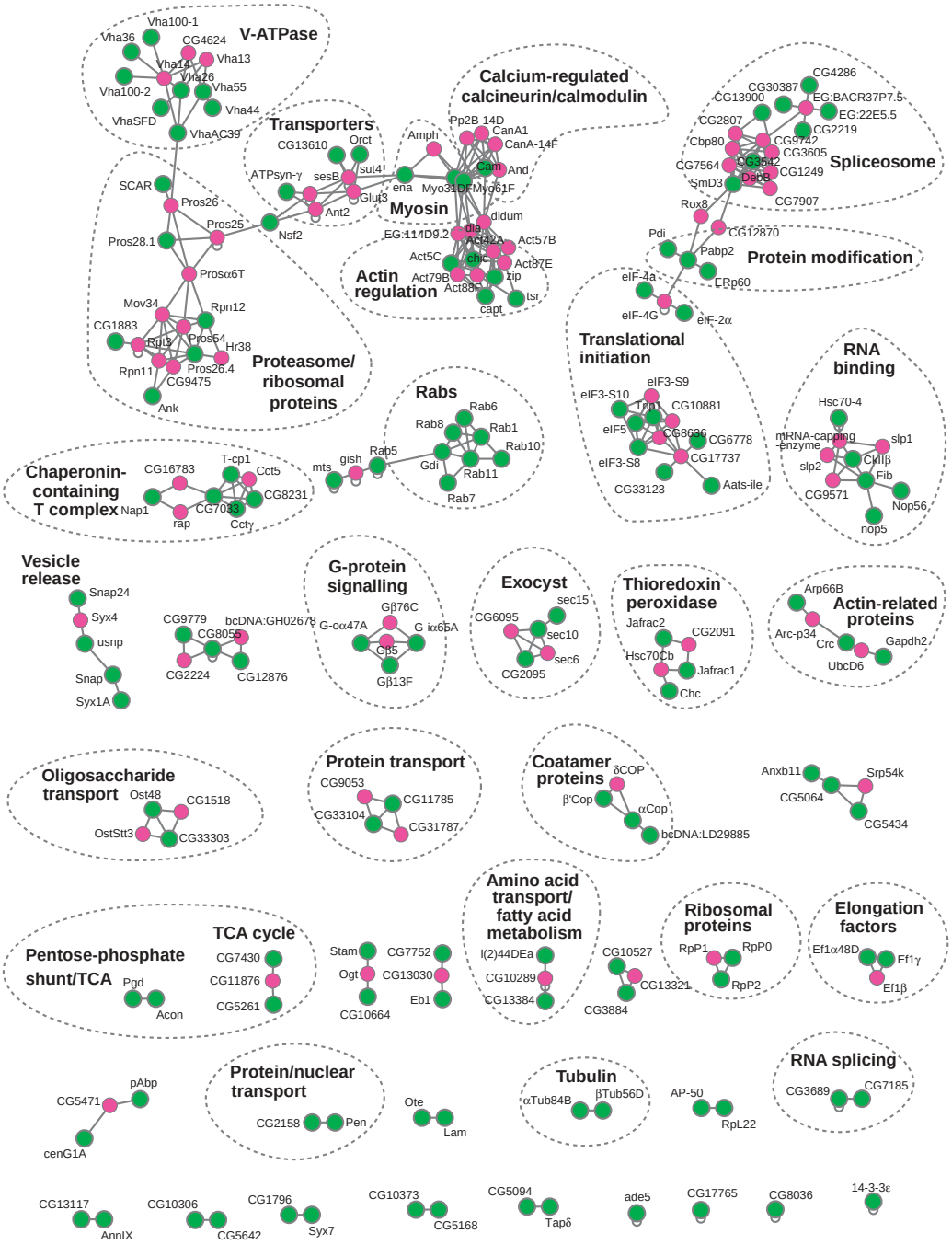


Figure 2 | Optimized phagosome networks. The interaction map was optimized (see Supplementary Methods) to define the most sensitive alternative. This map, corresponding to $(n,c) = (2,0.75)$, clearly defines clusters of functional modules and protein complexes. Green vertices are primary components and pink vertices are secondary components. TCA, tricarboxylic acid.

third, 100 genes decreased only *E. coli* internalization (53 with *z* scores of less than −1.5). Silencing of certain genes increased the uptake of bacteria by more than 50%, potentially representing negative regulators of phagocytosis (38 genes increased *S. aureus* uptake, 55 with *z* scores of more than 2.0; 34 genes increased *E. coli* internalization, 63 with *z* scores of more than 2.0; Fig. 3, and Supplementary Tables S6 and S7). Silencing of only one gene (*CG4046*, encoding a ribosomal protein) increased the uptake of both organisms.

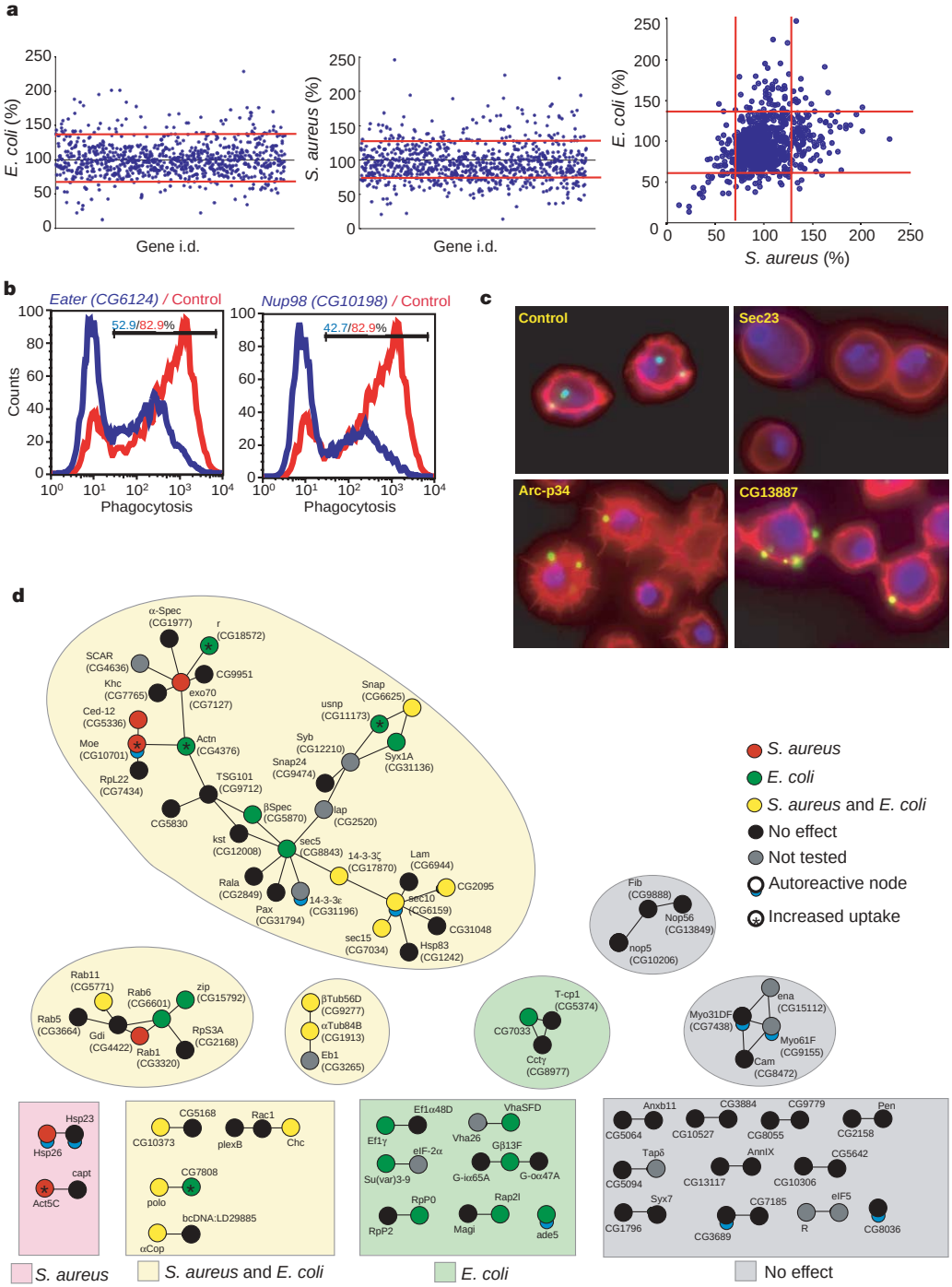


Figure 3 | Functional genomics to identify protein complexes and components involved in engulfment. **a**, The phagocytic index values after RNAi silencing of all phagosome genes on uptake of *S. aureus* and *E. coli*. Data are normalized to the mean of controls for each plate. **b**, Representative FACS plots of the phagocytosis assay demonstrating decreased *S. aureus* uptake after silencing of the two positive control genes, *CG6124* (*Eater*) and *CG10198* (*Nup98*) (blue histograms) compared with untreated controls (red histograms). **c**, Microscopy of changes in cell morphology and F-actin staining in cells with phagocytic defects after gene silencing. Silencing of Sec23 resulted in marked abnormality in actin polymerization and a total absence of bacterial binding; silencing of Arc-p34 (a member of the Arp2/3 complex) resulted in cells showing the formation of prominent filopodia and a failure to internalize bound bacteria; silencing of *CG13887* resulted in cells showing normal binding but failing to internalize bacteria. **d**, A protein–protein interaction network based on the high-confidence interactions between primary phagosome components was generated and the RNAi data were overlaid on this map to identify complexes involved in engulfment. Green nodes represent genes involved in *E. coli* uptake, red nodes represent genes involved in *S. aureus* uptake and yellow nodes are genes affecting uptake of both organisms. Unscreened genes are grey, and black nodes indicate genes with no effect on bacterial internalization. Complexes were arranged by effect on *S. aureus* (red areas), *E. coli* (green areas) and those regions affecting both organisms (yellow areas).

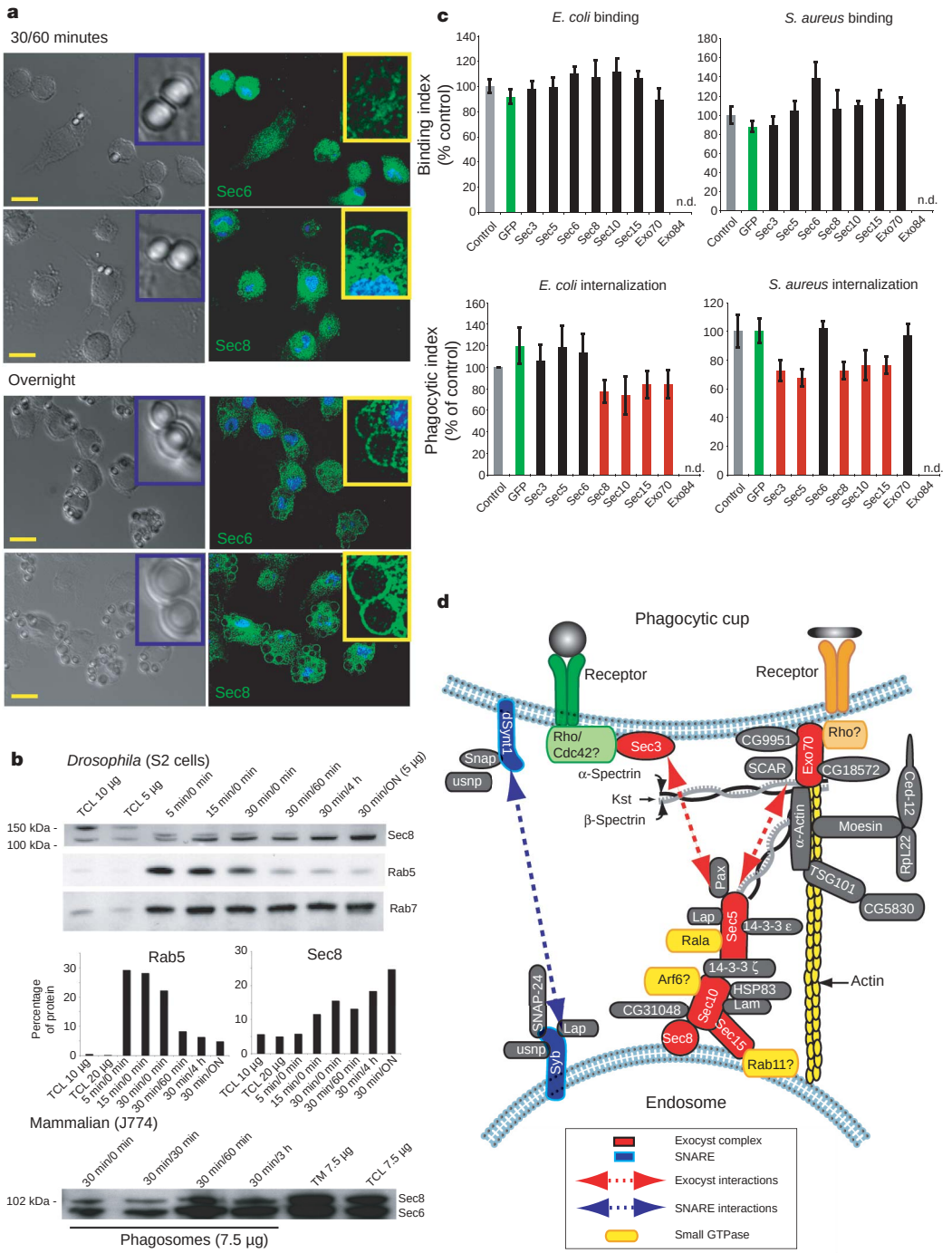


Figure 4 | Identification of the exocyst as a functional component of the phagosome. **a**, Fluorescence microscopy of Sec6/8 recruitment to mammalian 30 min/60 min (that is, 30 min binding on ice followed by 60 min of internalization at 37 °C) and late phagolysosomes. **b**, Western blots probed for the exocyst component Sec8 on isolated *Drosophila* phagosomes and Sec6 and Sec8 on mammalian phagosomes demonstrate the kinetics or recruitment of the exocyst to *Drosophila* and mammalian phagosomes. ON, overnight; TCL, total cell lysate; TM, total membrane. **c**, Effect of silencing exocyst components on bacterial binding and uptake. Red bars indicate RNAi treatments with a significant decrease compared with the irrelevant green fluorescent protein (GFP) control RNAi (green bar). n.d., not detected and hence not tested. Data are shown as means ± s.e.m. **d**, Model of the role of the exocyst derived from our combined iterative approach. Proteins were organized as predicted from the protein interaction data in Fig. 3d and oriented on the basis of the assumption that Exo70 and Sec3 would be recruited to the plasma membrane and Sec8–Sec10–Sec15 present on endosomes. These maps identify known exocyst interactors such as the small GTPase Rala and predicted previously unknown potential regulators and interactors of the exocyst. Integrating the proteomics with the RNAi data leads us to propose the following role for the exocyst during phagosome biogenesis. First, phagocytic receptors induce the differential activation of GTPases Rac, Rho and Cdc42 (for example the *Drosophila S. aureus* receptor is Rac2/RhoL dependent) to mediate engulfment. Second, depending on the ligands encountered and receptors engaged, the GTPase differentially recruits either Sec3 or Exo70 to the area of the plasma membrane from which the particle will be internalized. Third, Exo70 and/or Sec3 provide docking sites for the exocyst components Sec8, Sec10 and Sec15, which are present on the endosome membrane, thus facilitating the recruitment and tethering of endosomes to the phagocytic cup. Fourth, once endosomes and the phagocytic cup are in close proximity, SNAREs mediate membrane fusion. Exocyst components are shown in red.

In addition, to identify those protein modules that had a function in phagocytosis, the results of the RNAi screen were overlaid on the protein–protein interaction networks (Fig. 3d). This analysis identified the tubulin, Rab–GDI (Rab–GDP-dissociation inhibitor) and chaperonin-containing T-complexes as important for the internalization of bacteria (Fig. 3d). In addition the large cluster consisting of components of the fusome and exocyst also contained several molecules that affected phagocytosis. Certain complexes segregated into only one of the two pathogen-specific pathways (Fig. 3d). These data indicate that only certain phagosome components are involved in engulfment and suggest other functions for the remaining proteins associated with this organelle.

As an illustrative example we chose to focus on the exocyst[4]. The eight components of the exocyst complex (Sec3, Sec5, Sec6, Sec8, Sec10, Sec15, Exo70 and Exo84) can exist as separate proteins or as discrete subcomplexes or can assemble into an octodimeric complex. The exocyst assembles between the plasma membrane and secretory vesicles, tethering them before membrane fusion during exocytosis[21] and is also required for the transport of vesicles to lateral membranes in polarized cells, for the branching of neurites and for the formation of synapses. In addition to its role in exocytosis, exocyst components also concentrate in recycling endosomes and regulate receptor recycling and delivery of membrane to areas of localized plasma-membrane expansion[22,23]. Pertinent to phagocytosis, the exocyst components Sec10 and Sec15 interact directly with ADP-ribosylation factor 6 and Rab11, respectively, both of which regulate the recruitment of endosomes and membrane to the phagocytic cup[22,24–26]. Although the exocyst had hitherto not been implicated in phagocytosis, six of the eight known exocyst components were identified by proteomics and were also clustered within the interactome and by GoMiner ($P < 0.0001$).

Using immunofluoresence (Fig. 4a) and western blotting (Fig. 4b), we confirmed assembly of the exocyst on mammalian and *Drosophila* phagosomes, suggesting an evolutionarily conserved role for this complex in phagocytosis. We proposed that the exocyst might be involved either in the delivery of receptors during early phagocytosis or in membrane recruitment. Silencing of exocyst components did not decrease bacteria binding at 27 °C (conditions permissive for exocytosis (Fig. 4c), indicating that the exocyst was not required for receptor delivery. However, silencing of Sec8, Sec10 and Sec15 resulted in a 25–30% decrease in internalization (Fig. 4c), demonstrating a role for these components in the uptake of both *S. aureus* and *E. coli*. Sec3, Sec5 and Exo70 were differentially involved in uptake of the different organisms. These observations are consistent with observations that the exocyst acts not as a single macromolecular complex but assembles as distinct subcomplexes[27,28]. Specifically, Sec3 and Exo70 localize to target membranes, providing the docking site for other exocyst components[21] that are delivered from subcellular vesicles, such as endosomes, in which they concentrate[22,23]. It is possible that the involvement of Exo70 and Sec3 reflects differential receptor engagement by different pathogens. Unlike other regulators of early phagocytosis, Sec6 and Sec8 continued to accumulate on maturing phagosomes, with maximal association occurring after the loss of Rab5, a marker of immature phagosomes, and remained associated with latex-bead phagolysosomes for up to 16 h (Fig. 4a, b). These observations indicate that the exocyst might have additional roles in mature phagosomes. However, because silencing of the exocyst affected the efficient internalization of particles and the generation of phagosomes, we were unable to test this possibility definitively. These data lead us to propose a model (outlined in Fig. 4d) of a hitherto unknown but evolutionarily conserved role for the exocyst complex in phagocytosis. We suggest that receptor engagement (through the activation of small GTPases) recruits membrane exocyst components, Sec3 or Exo70, which act as docking localization sites on phagocytic cups. Subsequently, Sec8–Sec10–Sec15 assembles between endosomes and the phagocytic cup to facilitate membrane delivery and internalization. Additional roles of the exocyst on mature phagosomes remain to be defined.

The essential contribution of phagocytosis to innate and adaptive immunity underscores the importance of understanding the regulation of particle internalization and the organization of the phagosome. Using a combination of systems biology approaches we have generated a detailed model of the phagosome, identified novel regulators of phagocytosis and highlighted potentially unknown molecules and pathways involved in host defence. We suggest that some of these may be important, previously unidentified, targets for pathogens that evade host defence within the phagosome or disrupt functions of this organelle. Our 'systems-based model' has provided new insight into the functional organization of the phagosome and the necessary framework on which to build, in an iterative manner, to further our understanding of phagocytosis. We provide evidence that this model can be extended to mammalian phagocytosis, adding new dimensions to our understanding of the host–pathogen interaction and other aspects of innate and adaptive immunity.

METHODS

Phagosome isolation, proteomics and protein identification were performed as described previously[29]; and detailed methods are available in Supplementary Methods. Generation of the phagosome interactome and network inference and analysis were performed as described in the text and in Supplementary Methods. Double-stranded RNA was generated from a *Drosophila* RNAi library containing 13,000 genes (Geneservice) and the RNAi screen was performed as described previously[20]. Primer sequences used to generate each double-stranded RNA can be located using the MRCg plate and well identifications in Supplementary Table S8 and the Geneservice website (http://www.geneservice.co.uk/home/) (further information is available in Supplementary Methods).

GO Miner. Gene ontology classifications were drawn from the Gene Ontology website (http://www.geneontology.org). The enrichment factor (*Re*) found in the DAG (directed acyclic graph) files and the *P* value for statistical significance were calculated by GO Miner as follows. The enrichment factor *Re* is defined as $(n_f/n)/(N_f/N)$, where n_f is phagosome proteins in a specific category, n is the total number of phagosome proteins, N_f is the total number of proteins in a specific category in the representative data set, and N is the total number of proteins in the representative data set. GO Miner uses two-sided Fisher's exact test to determine the *P* value for a category and tests the null hypothesis that the category is neither enriched in, nor depleted of, flagged genes with regard to what would have been expected by chance alone. Full data sets used to derive these data are available on request. In addition, GO Miner generated data on the enrichment of BioCarta pathways, details of which are available in Supplementary Table S5 or on the BioCarta website (http://www.biocarta.com).

Received 13 July; accepted 24 October 2006.
Published online 6 December 2006.

1. Greenberg, S. & Grinstein, S. Phagocytosis and innate immunity. *Curr. Opin. Immunol.* **14,** 136–145 (2002).
2. Aderem, A. & Underhill, D. M. Mechanisms of phagocytosis in macrophages. *Annu. Rev. Immunol.* **17,** 593–623 (1999).
3. Stuart, L. M. & Ezekowitz, R. A. Phagocytosis: elegant complexity. *Immunity* **22,** 539–550 (2005).
4. TerBush, D. R., Maurice, T., Roth, D. & Novick, P. The exocyst is a multiprotein complex required for exocytosis in *Saccharomyces cerevisiae. EMBO J.* **15,** 6483–6494 (1996).
5. Desjardins, M. ER-mediated phagocytosis: a new membrane for new functions. *Nature Rev. Immunol.* **3,** 280–291 (2003).
6. Desjardins, M. & Griffiths, G. Phagocytosis: latex leads the way. *Curr. Opin. Cell Biol.* **15,** 498–503 (2003).
7. Garin, J. *et al.* The phagosome proteome: insight into phagosome functions. *J. Cell Biol.* **152,** 165–180 (2001).
8. Pearson, A. M. *et al.* Identification of cytoskeletal regulatory proteins required for efficient phagocytosis in *Drosophila. Microbes Infect.* **5,** 815–824 (2003).
9. Ramet, M. *et al. Drosophila* scavenger receptor CI is a pattern recognition receptor for bacteria. *Immunity* **15,** 1027–1038 (2001).
10. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19,** 551–560 (2003).
11. Vidal, M. Interactome modeling. *FEBS Lett.* **579,** 1834–1838 (2005).
12. Bader, J. S. Greedily building protein networks with confidence. *Bioinformatics* **19,** 1869–1874 (2003).
13. Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnol.* **22,** 78–85 (2004).

14. Uetz, P. & Finley, R. L. Jr. From protein networks to biological systems. *FEBS Lett.* **579,** 1821–1827 (2005).
15. Matthews, L. R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.* **11,** 2120–2126 (2001).
16. Asthana, S., King, O. D., Gibbons, F. D. & Roth, F. P. Predicting protein complex membership using probabilistic network reliability. *Genome Res.* **14,** 1170–1175 (2004).
17. Cox, D. *et al.* Myosin X is a downstream effector of PI(3)K during phagocytosis. *Nature Cell Biol.* **4,** 469–477 (2002).
18. Gold, E. S. *et al.* Amphiphysin IIm, a novel amphiphysin II isoform, is required for macrophage phagocytosis. *Immunity* **12,** 285–292 (2000).
19. Colucci-Guyon, E. *et al.* A role for mammalian diaphanous-related formins in complement receptor (CR3)-mediated phagocytosis in macrophages. *Curr. Biol.* **15,** 2007–2012 (2005).
20. Ramet, M., Manfruelli, P., Pearson, A., Mathey-Prevot, B. & Ezekowitz, R. A. Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli*. *Nature* **416,** 644–648 (2002).
21. Boyd, C., Hughes, T., Pypaert, M. & Novick, P. Vesicles carry most exocyst subunits to exocytic sites marked by the remaining two subunits, Sec3p and Exo70p. *J. Cell Biol.* **167,** 889–901 (2004).
22. Zhang, X. M., Ellis, S., Sriratana, A., Mitchell, C. A. & Rowe, T. Sec15 is an effector for the Rab11 GTPase in mammalian cells. *J. Biol. Chem.* **279,** 43027–43034 (2004).
23. Prigent, M. *et al.* ARF6 controls post-endocytic recycling through its downstream exocyst complex effector. *J. Cell Biol.* **163,** 1111–1121 (2003).
24. Bajno, L. *et al.* Focal exocytosis of VAMP3-containing vesicles at sites of phagosome formation. *J. Cell Biol.* **149,** 697–706 (2000).
25. Niedergang, F., Colucci-Guyon, E., Dubois, T., Raposo, G. & Chavrier, P. ADP ribosylation factor 6 is activated and controls membrane delivery during phagocytosis in macrophages. *J. Cell Biol.* **161,** 1143–1150 (2003).
26. Cox, D., Lee, D. J., Dale, B. M., Calafat, J. & Greenberg, S. A. Rab11-containing rapidly recycling compartment in macrophages that promotes phagocytosis. *Proc. Natl Acad. Sci. USA* **97,** 680–685 (2000).
27. Clandinin, T. R. Surprising twists to exocyst function. *Neuron* **46,** 164–166 (2005).
28. Mehta, S. Q. *et al.* Mutations in *Drosophila* sec15 reveal a function in neuronal targeting for a subset of exocyst components. *Neuron* **46,** 219–232 (2005).
29. Desjardins, M., Huber, L. A., Parton, R. G. & Griffiths, G. Biogenesis of phagolysosomes proceeds through a sequential series of interactions with the endocytic apparatus. *J. Cell Biol.* **124,** 677–688 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Kearney and J. Bergeron from the Montreal Proteomics Network, and Genome-Quebec-Canada for their support. J.S.B., M.D. and R.A.B.E. thank their laboratories for their support. The work was supported by a Wellcome Trust Clinician Scientist Award to L.M.S., grants from the Whitaker Foundation and NIH/NIGMS to J.S.B, grants from the Canadian Institute for Health Research and Genome-Canada-Québec to M.D., and NIH grants to R.A.B.E. The work was conceived through discussions between the Laboratory of Developmental Immunology, Massachusetts General Hospital/Harvard Medical School and the Département de pathologie et biologie cellulaire, Université de Montréal. The bioinformatics and RNAi screens were performed in the Laboratory of Developmental Immunology, Massachusetts General Hospital/Harvard Medical School; the protein–protein networks were generated in the Department of Biomedical Engineering and High-Throughput Biology Center, Johns Hopkins University; the proteomics, the annotation of the components and the phagosome isolation were performed in the Département de pathologie et biologie cellulaire, Université de Montréal.

Author Contributions L.M.S. and J.B. contributed equally to this work. J.S.B., M.D. and R.A.B.E. contributed equally to this work. The manuscript was written by L.M.S. and the website linked to this paper was designed by G.M.C.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.M.S. (lstuart@partners.org).

LETTERS

Hsp70 regulates erythropoiesis by preventing caspase-3-mediated cleavage of GATA-1

Jean-Antoine Ribeil[1]*, Yael Zermati[1,2]*, Julie Vandekerckhove[1], Severine Cathelin[2], Joelle Kersual[1], Michaël Dussiot[1], Séverine Coulon[1], Ivan Cruz Moura[1], Ann Zeuner[3], Thomas Kirkegaard-Sørensen[4], Bruno Varet[1,5], Eric Solary[2], Carmen Garrido[2] & Olivier Hermine[1,5]

Caspase-3 is activated during both terminal differentiation and erythropoietin-starvation-induced apoptosis of human erythroid precursors. The transcription factor GATA-1, which performs an essential function in erythroid differentiation[1,2] by positively regulating promoters of erythroid and anti-apoptotic genes[3–6], is cleaved by caspases in erythroid precursors undergoing cell death upon erythropoietin starvation or engagement of the death receptor Fas[7,8]. In contrast, by an unknown mechanism, GATA-1 remains uncleaved when these cells undergo terminal differentiation upon stimulation with Epo[9–11]. Here we show that during differentiation, but not during apoptosis, the chaperone protein Hsp70 protects GATA-1 from caspase-mediated proteolysis. At the onset of caspase activation, Hsp70 co-localizes and interacts with GATA-1 in the nucleus of erythroid precursors undergoing terminal differentiation. In contrast, erythropoietin starvation induces the nuclear export of Hsp70 and the cleavage of GATA-1. In an *in vitro* assay, Hsp70 protects GATA-1 from caspase-3-mediated proteolysis through its peptide-binding domain. The use of RNA-mediated interference to decrease the Hsp70 content of erythroid precursors cultured in the presence of erythropoietin leads to GATA-1 cleavage, a decrease in haemoglobin content, downregulation of the expression of the anti-apoptotic protein Bcl-X_L, and cell death by apoptosis. These effects are abrogated by the transduction of a caspase-resistant GATA-1 mutant. Thus, in erythroid precursors undergoing terminal differentiation, Hsp70 prevents active caspase-3 from cleaving GATA-1 and inducing apoptosis.

During erythropoiesis, caspase activation is required for maturation of erythroblasts. In this process, however, in contrast to what occurs during the apoptosis of erythroblasts, some targets, including GATA-1, remain uncleaved (Supplementary Fig. S1). The fate of erythroblasts is therefore determined downstream of caspase activation by an unknown mechanism.

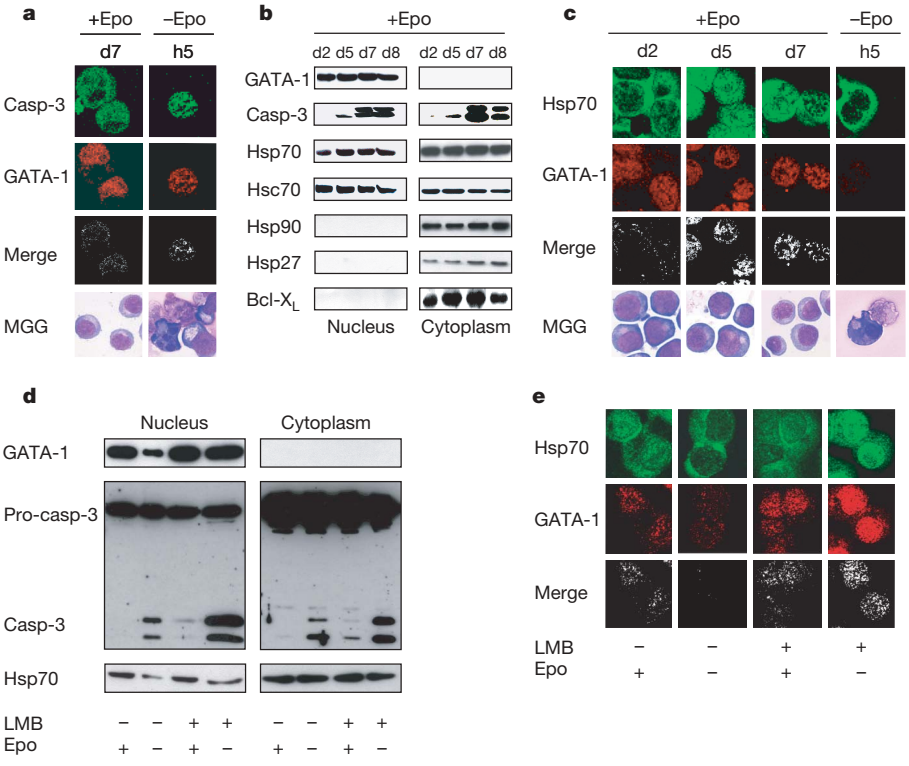


Figure 1 | Hsp70 nuclear expression and co-localization with GATA-1 in differentiating erythroblasts is lost during erythropoietin-starvation-induced apoptosis. **a**, Top three rows: co-localization (white) of GATA-1 (red) and active caspase-3 (p17 subunit, green) at day 7 with erythropoietin (+Epo, d7) or after cytokine starvation for 5 h (−Epo, h5). Bottom row: morphological (MGG) analysis of cytokine-deprived cells (magnification ×40). **b**, Protein expression for the indicated days (d2 to d8) in the presence of erythropoietin. **c**, Top three rows: Hsp70 (green) and GATA-1 (red) co-localize in the nucleus in the presence of erythropoietin. Bottom row: MGG analysis ($n = 5$). **d**, **e**, Immunoblot analysis (**d**) and confocal microscopy analysis (**e**) at day 4, pretreated (+) or not (−) with 20 nM leptomycin B (LMB) for 1 h before being starved of erythropoietin (−Epo) or not (+Epo) for 5 h. One representative experiment of three is shown. Casp, caspase.

[1]CNRS UMR 8147, Faculté de Médecine et Université René Descartes Paris V, Institut Fédérative Necker, 75270 Paris, France. [2]INSERM UMR 517, 21079 Dijon, France. [3]Department of Hematology and Oncology, Istituto Superiore di Santia, 00161 Roma, Italy. [4]Department of Apoptosis, Institute for Cancer Biology, Danish Cancer Society, Strandboulevarden 49, 2100 Copenhagen. [5]Department of Hematology, Faculté de Médecine et Université René Descartes Paris V, Assistance publique des hôpitaux de Paris, Necker, 75270 Paris, France.
*These authors contributed equally to this work.

During erythroid differentiation, caspase-3, and to a smaller extent caspase-7 but not other caspases (Supplementary Fig. S2), is found activated in the nucleus and co-localized with GATA-1 (Fig. 1a, b). Bcl-X_L, a key survival factor in erythropoiesis, prevents the release of pro-apoptotic molecules from the mitochondria to the cytosol, upstream of caspase activation[12,13]; it therefore may not account for the differential effect of activated caspases in erythropoietin-stimulated and erythropoietin-deprived erythroblasts. Other protective proteins include stress-inducible proteins such as the heat-shock proteins Hsp90, Hsp70 and Hsp27. These molecular chaperones modulate the assembly, transport and folding of other proteins[14]. They also have key functions in cell survival after stressful stimuli, for example by associating with apoptotogenic proteins downstream of the mitochondria, including cytochrome *c* (ref. 15), Apaf-1 (ref. 16) and apoptosis-inducing factor (AIF)[17]. The inducible Hsp70 was shown also to rescue cells from apoptosis at a later stage than any other known survival factor, downstream of caspase activation[18]. Although its role is unknown, Hsp70 is expressed constitutively in erythroid cells of *Xenopus laevis*[19]. On the basis of these observations, we examined whether Hsp70 was expressed during human erythroid differentiation and whether it could protect GATA-1 from cleavage by activated caspases. We studied Hsp70 expression, localization and interaction with GATA-1 during both erythroblast differentiation and erythropoietin-starvation-induced apoptosis.

Hsp27, Hsp70 and Hsp90 proteins were expressed constitutively in human erythroblasts undergoing differentiation (Fig. 1b). Of these three proteins, only Hsp70 was highly expressed in the nucleus of differentiating cells (Fig. 1b), where it co-localized with GATA-1 (Fig. 1c). Similar results were found in fresh bone marrow glycophorin-A-positive cells (Supplementary Fig. S3), excluding the possibility that the nuclear expression of Hsp70 was the consequence of cell culture. The intensity of GATA-1 and Hsp70 co-localization increased with the level of caspase activation (Fig. 1b, c). In contrast, during erythropoietin-starvation-induced apoptosis, Hsp70 lost nuclear localization, which correlated with a decreased expression of GATA-1 (Fig. 1c–e), indicating caspase-3-mediated cleavage[8]. Addition of the Crm1-mediated nuclear export inhibitor leptomycin B to erythropoietin-starved cells prevented the nuclear export of Hsp70 and degradation of GATA-1, despite a higher level of activated caspase-3 than observed during apoptosis (Fig. 1d, e). Moreover, in the presence of leptomycin B, Hsp70 co-localized with GATA-1 in the nucleus of erythropoietin-starved erythroblasts (Fig. 1d, e).

Immunoprecipitation experiments on whole-cell extracts (Fig. 2a) as well as nuclear extracts (not shown) from differentiated erythroblasts demonstrated that Hsp70 co-immunoprecipated with GATA-1, whereas neither Hsp90 nor Hsp27 interacted with the transcription factor. This interaction seemed to be specific because Hsp70 did not interact with lamin B, another nuclear protein that is cleaved by caspases in differentiating erythroid cells[9], or with acinus (data not shown). By using Hsp70 deletion mutants, we observed that binding of Hsp70 to GATA-1 involved the peptide-binding domain of Hsp70 (Fig. 2b, c). After withdrawal of erythropoietin, co-immunoprecipitation between Hsp70 and GATA-1 was decreased as a consequence of GATA-1 cleavage, but the remaining Hsp70 in the nucleus still interfered with and protected GATA-1, as demonstrated by increasing GATA-1 input (Supplementary Fig. S4). Taken together, these data indicate that the peptide-binding domain of Hsp70 was required to maintain GATA-1 integrity when caspases were activated during erythroid differentiation.

Accordingly, an *in vitro* proteolysis assay showed that recombinant Hsp70 protected GATA-1 from cleavage by caspase-3 in a dose-dependent manner, whereas recombinant Hsp90 did not (Fig. 2d). The Hsp70-mediated protection seemed to be protein-specific because, in this *in vitro* assay, Hsp70 did not prevent the caspase-3-induced cleavage of poly(ADP-ribose) polymerase, a protein cleaved in erythroid cells that are undergoing differentiation (Fig. 2d). In accordance with co-immunoprecipitation experiments, the GATA-1 protection was lost when the peptide-binding domain of Hsp70 was deleted (Fig. 2e).

To further show the role of Hsp70 in determining the fate of erythroblasts, Hsp70 expression was inhibited with an approach involving RNA-mediated interference. Transient transfection of Hsp70-specific short interfering RNAs (siRNA Hsp70) in erythroid precursors significantly decreased the Hsp70 protein level 7.5-fold and 4-fold at days 2 and 4, respectively (Fig. 3a, b); this decrease was observed both in the nucleus and in the cytoplasm (Supplementary Fig. S5). At day 6, Hsp70 level in Hsp70-specific siRNA-transfected cells had returned to the level observed in scramble siRNA-transfected erythroid cells (data not shown). To achieve a greater inhibition of Hsp70 at the time of onset of caspase-3 activation, terminal erythroid differentiation was accelerated by omitting stem cell factor (SCF) from the culture medium[20], because this omission did not increase cell apoptosis (data not shown). In these culture conditions, caspase-3 remained inactivated at day 2 and no significant difference was observed between Hsp70-specific and control siRNA-transfected cells with regard to GATA-1 expression, apoptosis and cell differentiation. At the onset of caspase-3 activation (day 4), GATA-1 remained intact in control siRNA-transfected cells, whereas the protein was almost completely degraded in Hsp70-depleted cells (Fig. 3a, b). As

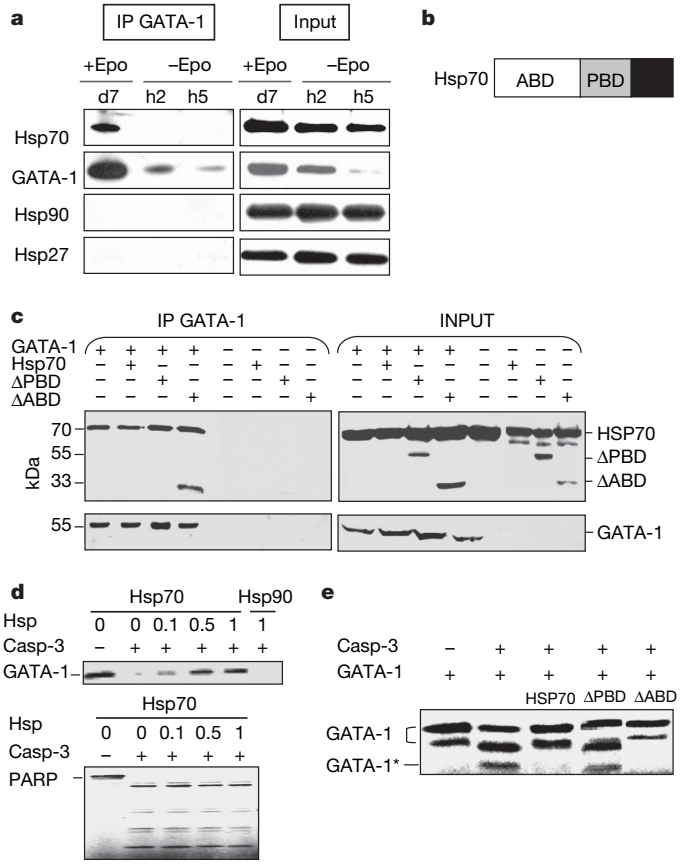


Figure 2 | Co-immunoprecipitation of GATA-1 and Hsp70 during erythroid differentiation. **a**, Immunoprecipitation (IP) of GATA-1 in whole cell extracts at day 7 with erythropoietin (+Epo, d7) or after cytokine starvation for 2 and 5 h (−Epo, h2 and h5), and immunoblotting for indicated proteins. **b**, Representation of Hsp70: ATP-binding domain (ABD) and peptide-binding domain (PBD). **c**, Immunoprecipitation of GATA-1 in GATA-1-transduced HeLa cells transfected with plasmids encoding Hsp70 or mutant (ΔABD, ΔPBD) proteins. **d**, **e**, SDS–PAGE analysis of *in vitro*-translated ^{35}S-labelled GATA-1 and poly(ADP-ribose) polymerase (PARP) exposed to recombinant caspase-3 and increasing amounts (0–1 μg) of recombinant Hsp70 or Hsp90 (1 μg) (**d**) or full-length or deleted mutants of Hsp70 (**e**). GATA-1* indicates a GATA-1 cleaved fragment. In all panels, one representative experiment of three is shown. Casp, caspase.

expected, the degradation of GATA-1 in Hsp70-depleted erythroblasts at day 4 was associated with a strong decrease in the expression of GATA-1-regulated gene products; for example, fewer than 10% of cells were haemoglobinized, in comparison with almost 50% in control cells ($P < 0.005$; data not shown) and Bcl-X_L protein level was decreased (Fig. 3b). Depletion of Hsp70 also markedly decreased the total cell number (Fig. 3d) while increasing the rate of cell death at day 4 (40% in Hsp70-depleted versus 7% in control siRNA-transfected cells; $P = 0.007$; Fig. 3e). Morphological analysis of cells indicated that Hsp70 depletion induced a decrease in cell differentiation (Fig. 3c, f); that is, the proportion of immature erythroblasts (identified by their basophilic cytoplasm and larger size) with nuclear chromatin condensation was only 10% in Hsp70-depleted cells at day 4, which was similar to that in control cells, whereas the proportion of mature erythroblasts (identified by their acidophilic cytoplasm and smaller size) exhibiting nuclear features of apoptosis was about 90% in Hsp70-depleted cells ($P = 0.0004$) (Fig. 3c, g).

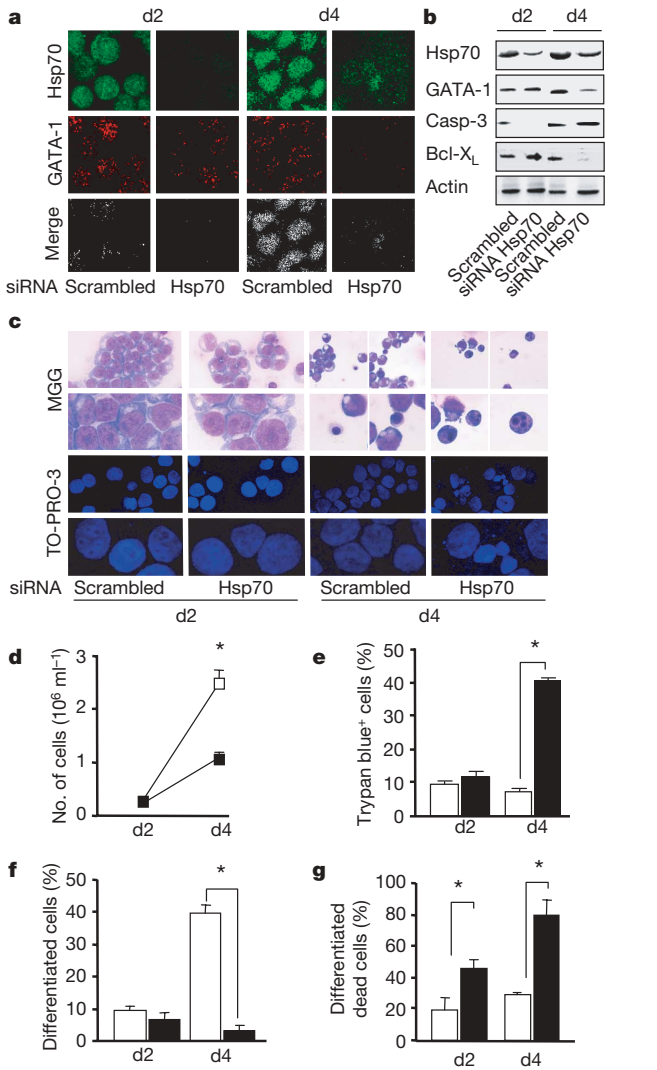


Figure 3 | Decreased Hsp70 content induces death of erythroblasts undergoing differentiation. Analysis of erythroid progenitors 2 days (d2) and 4 days (d4) after transfection with siRNA targeting Hsp70 or a scrambled siRNA. **a**, Confocal microscopy analysis of GATA-1 and Hsp70 expression. **b**, Immunoblot analysis of indicated proteins in whole-cell lysates. **c**, Cell and nucleus morphology assessed by MGG and TO-PRO-3 analysis at various magnifications. **d**, Growth curves. **e**, Percentage of cell death. **f**, Percentage of mature cells. **g**, Percentage of mature cells exhibiting morphological features of apoptosis. Graphed results are means and s.e.m. for three independent experiments; open bars and symbols, scrambled siRNA; filled bars and symbols, Hsp70 siRNA. Asterisk, $P < 0.05$.

Similar results were obtained when Hsp70 expression was decreased by the use of antisense oligonucleotides (Supplementary Fig. S6). We also observed the same results in the presence of SCF, by using a stealth Hsp70 siRNA that exhibits a longer half-life than standard siRNAs, ruling out an increase in apoptosis due to the lack of SCF in the culture medium (Supplementary Fig. S7). Taken together, these findings indicated that Hsp70 depletion prevented erythroid cells from undergoing terminal differentiation by allowing activated caspase-3 to trigger apoptotic cell death.

To check whether the effect of Hsp70-targeting siRNA was due to GATA-1 cleavage rather than to a general effect on cellular metabolism, a previously described GATA-1 mutant (μG1)[8] that resists caspase-mediated cleavage was retrovirally transduced in CD34-positive cells. The same vector encoding wild-type GATA-1 (WT G1) and the empty vector were used as controls. Cell proliferation was decreased by 50% both in cells transfected with wild-type G1 and cells in transfected with μG1, in accordance with previous observations[21]. As expected, transfection with siRNA targeting Hsp70 in cells transduced with either the empty or the wild-type GATA-1-encoding vectors resulted in a decrease in GATA-1 expression (Fig. 4a), cell expansion (Fig. 4b) and cell maturation (Fig. 4d), while increasing

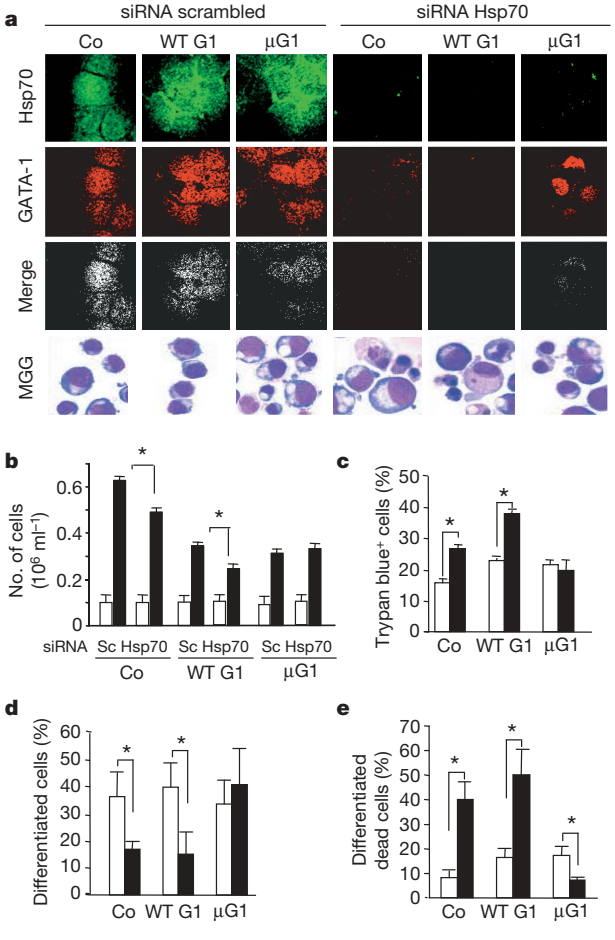


Figure 4 | Transduction of caspase-resistant GATA-1 mutant protects erythroblasts from death after depletion of Hsp70. $CD34^+$ haematopoietic cells transduced with an empty vector (Co), a vector encoding wild-type GATA-1 (WT G1), or Asp 125-mutated GATA-1 (μG1) were transfected with Hsp70 siRNA or scrambled siRNA (Sc). **a**, Confocal microscopy analysis of GATA-1 and Hsp70 expression. Cell maturation was assessed by morphological (MGG) analysis. **b**, Cell expansion assessed at day 2 (open bars) or day 4 (filled bars). **c**, Percentage of trypan-blue-positive cells. **d**, Percentage of differentiated cells. **e**, Percentage of differentiated cells exhibiting morphological features of apoptosis all at day 4. In **c**–**e**, open bars, scrambled siRNA; filled bars, Hsp70 siRNA. Graphed results are means and s.e.m. for three independent experiments. Asterisk, $P < 0.05$.

mature cell death (Fig. 4c, e). In contrast, Hsp70 siRNA had no significant effect on μGATA-1 transduced cells (Fig. 4). Taken together, these results suggested that Hsp70 protected differentiating erythroblasts from apoptosis through the inhibition of GATA-1 cleavage by caspase-3.

On the basis of these observations, we propose a model in which erythropoietin protects erythroid cells undergoing differentiation from caspase-mediated apoptosis by regulating the cellular localization of Hsp70 (Supplementary Fig. S8). This model provides a potential explanation for the lack of cell death when caspases are activated in erythroid cells undergoing terminal differentiation[9–11]. Bcl-X_L is a key protein in erythroid progenitor survival[4]; it acts by preventing the release of pro-apoptotic molecules from the mitochondria[7,22]. Our results strongly indicate that Hsp70 is another key erythroid antiapoptotic protein that acts both upstream and downstream of Bcl-X_L. First, by protecting GATA-1 from caspase-3-mediated cleavage, Hsp70 maintains Bcl-X_L expression, in synergy with erythropoietin. Second, Hsp70 saves erythroid progenitors in which caspases are activated from apoptosis-inducing GATA-1 proteolysis. We cannot exclude the possibility that, in addition to these two effects, Hsp70 limits caspase-3 activation by interacting with Apaf-1, thus preventing the formation of the apoptosome[16,23].

Thus, our data indicate that the fate of erythroblasts—apoptosis versus differentiation—is determined downstream of caspase activation by the nuclear localization of Hsp70.

METHODS

Erythroid cells were generated as described previously[24]. Details of reagents and protocols for cell proliferation and differentiation analysis, immunoblot analysis, immunoprecipitation, transfection of siRNAs or antisense cDNA targeting Hsp70 in erythroid progenitors, the production of retroviral particles and the infection of haematopoietic progenitors, the generation of Hsp70 mutants, *in vitro* GATA-1 cleavage assays and confocal fluorescence microscopy analysis are provided in Supplementary Methods. Statistical analyses were performed with the Statview software package. Comparisons were made by analysis of variance. Data are expressed as means and s.e.m. Differences were considered significant at $P < 0.05$.

Received 16 June; accepted 25 October 2006.
Published online 10 December 2006.

1. Fujiwara, Y. *et al.* Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc. Natl Acad. Sci. USA* **93,** 12355–12358 (1996).
2. Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349,** 257–260 (1991).
3. Weiss, M. J., Keller, G. & Orkin, S. H. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1 embryonic stem cells. *Genes Dev.* **8,** 1184–1197 (1994).
4. Motoyama, N. *et al. bcl-x* prevents apoptotic cell death of both primitive and definitive erythrocytes at the end of maturation. *J. Exp. Med.* **189,** 1691–1698 (1999).
5. Weiss, M. J. & Orkin, S. H. Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis. *Proc. Natl Acad. Sci. USA* **92,** 9623–9627 (1995).
6. Gregory, T. *et al.* GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-x_L expression. *Blood* **94,** 87–96 (1999).
7. Gregoli, P. A. & Bondurant, M. C. Function of caspases in regulating apoptosis caused by erythropoietin deprivation in erythroid progenitors. *J. Cell. Physiol.* **178,** 133–143 (1999).
8. De Maria, R. *et al.* Negative regulation of erythropoiesis by caspase-mediated cleavage of GATA-1. *Nature* **401,** 489–493 (1999).
9. Zermati, Y. *et al.* Caspase activation is required for terminal erythroid differentiation. *J. Exp. Med.* **193,** 247–254 (2001).
10. Carlile, G. W., Smith, D. H. & Wiedmann, M. Caspase-3 has a nonapoptotic function in erythroid maturation. *Blood* **103,** 4310–4316 (2004).
11. Kolbus, A. *et al.* Raf-1 antagonizes erythroid differentiation by restraining caspase activation. *J. Exp. Med.* **196,** 1347–1353 (2002).
12. Shimizu, S., Narita, M. & Tsujimoto, Y. Bcl-2 family proteins regulate the release of apoptogenic cytochrome *c* by the mitochondrial channel VDAC. *Nature* **399,** 483–487 (1999).
13. Adams, J. M. & Cory, S. The Bcl-2 protein family: arbiters of cell survival. *Science* **281,** 1322–1326 (1998).
14. Garrido, C. *et al.* HSP27 and HSP70: potentially oncogenic apoptosis inhibitors. *Cell Cycle* **2,** 579–584 (2003).
15. Bruey, J. M. *et al.* Hsp27 negatively regulates cell death by interacting with cytochrome *c*. *Nature Cell Biol.* **2,** 645–652 (2000).
16. Beere, H. M. *et al.* Heat-shock protein 70 inhibits apoptosis by preventing recruitment of procaspase-9 to the Apaf-1 apoptosome. *Nature Cell Biol.* **2,** 469–475 (2000).
17. Ravagnan, L. *et al.* Heat-shock protein 70 antagonizes apoptosis-inducing factor. *Nature Cell Biol.* **3,** 839–843 (2001).
18. Jaattela, M. *et al.* Hsp70 exerts its anti-apoptotic function downstream of caspase-3-like proteases. *EMBO J.* **17,** 6124–6134 (1998).
19. Winning, R. S. & Browder, L. W. Changes in heat shock protein synthesis and hsp70 gene transcription during erythropoiesis of *Xenopus laevis*. *Dev. Biol.* **128,** 111–120 (1988).
20. Muta, K. *et al.* Stem cell factor retards differentiation of normal human erythroid progenitor cells while stimulating proliferation. *Blood* **86,** 572–580 (1995).
21. Munugalavadla, V. *et al.* Repression of c-kit and its downstream substrates by GATA-1 inhibits cell proliferation during erythroid maturation. *Mol. Cell. Biol.* **25,** 6747–6759 (2005).
22. Gregoli, P. A. & Bondurant, M. C. The roles of Bcl-X_L and apopain in the control of erythropoiesis by erythropoietin. *Blood* **90,** 630–640 (1997).
23. Saleh, A. *et al.* Negative regulation of the Apaf-1 apoptosome by Hsp70. *Nature Cell Biol.* **2,** 476–483 (2000).
24. Zermati, Y. *et al.* Transforming growth factor inhibits erythropoiesis by blocking proliferation and accelerating differentiation of erythroid progenitors. *Exp. Hematol.* **28,** 885–894 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Pouzet for her assistance in confocal analysis, F. Valensi and V. Asnafi for their assistance in cytological analysis, Y. Dumez, A. Benachi and F. Audat for providing us with cord blood samples; U. Testa for the cDNAs of GATA-1 and poly(ADP-ribose) polymerase subcloned in PET21; and A. Benmerah for providing us with leptomycin B. This work was supported by grants from the Ligue nationale contre le cancer (LNC), the Fondation pour la recherche médicale (FRM), the Association pour la recherche sur le cancer (ARC), Cancéropole d'Île de France, Fondation de France, Ministère de la recherche and AMGEN.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to O.H. (hermine@necker.fr) or Y.Z. (zermati@igr.fr).

LETTERS

A human colon cancer cell capable of initiating tumour growth in immunodeficient mice

Catherine A. O'Brien[1], Aaron Pollett[2], Steven Gallinger[3] & John E. Dick[1,4]

Colon cancer is one of the best-understood neoplasms from a genetic perspective[1–3], yet it remains the second most common cause of cancer-related death, indicating that some of its cancer cells are not eradicated by current therapies[4,5]. What has yet to be established is whether every colon cancer cell possesses the potential to initiate and sustain tumour growth, or whether the tumour is hierarchically organized so that only a subset of cells—cancer stem cells—possess such potential[6,7]. Here we use renal capsule transplantation in immunodeficient NOD/SCID mice to identify a human colon cancer-initiating cell (CC-IC). Purification experiments established that all CC-ICs were $CD133^+$; the $CD133^-$ cells that comprised the majority of the tumour were unable to initiate tumour growth. We calculated by limiting dilution analysis that there was one CC-IC in 5.7×10^4 unfractionated tumour cells, whereas there was one CC-IC in 262 $CD133^+$ cells, representing >200-fold enrichment. CC-ICs within the $CD133^+$ population were able to maintain themselves as well as differentiate and re-establish tumour heterogeneity upon serial transplantation. The identification of colon cancer stem cells that are distinct from the bulk tumour cells provides strong support for the hierarchical organization of human colon cancer, and their existence suggests that for therapeutic strategies to be effective, they must target the cancer stem cells.

Human tumour biology has long been studied in experimental xenogeneic colon cancer models, typically generated by injecting cell lines or implanting pieces of primary tumours into immunodeficient mice[8–10]. However, cell lines do not recapitulate all aspects of primary tumours and a quantitative assay for single cells is required to determine whether CC-ICs exist in colon cancer. Therefore, we developed a reliable xenograft model through subrenal capsule implantation of human colon cancer cell suspensions into pre-irradiated non-obese diabetic (NOD)/severe-combined immunodeficient (SCID) mice. Tumour formation occurred in 17 out of 17 samples tested, comprising six primary colon cancers, ten liver metastases, and one retroperitoneal metastasis (Table 1 and Supplementary Fig. 1a and b). The histology and degree of differentiation of all xenografts resembled the original tumours from which they were derived (Fig. 1). The tumours were positive for cytokeratin-20 (CK-20) and negative for cytokeratin-7 (CK-7), a pattern seen almost exclusively in colonic adenocarcinoma[11]. Xenografts and parent tumours exhibited similar patterns of expression for multiple mucin antigens and for markers highly associated with colon cancers including carcinoembryonic antigen (CEA)[12] and p53 (ref. 13). The degree of tumour cell proliferation, as revealed by MIB-1 staining[12], was similar in xenografts and parent tumours (Fig. 1). Thus, the xenografts generated in this model matched the phenotypes of the original tumours.

To determine whether this xenotransplant system was quantitative and able to detect single CC-ICs, we performed limiting dilution experiments. Groups of NOD/SCID mice were transplanted with replicate doses of human colon cancer cells over a range from doses unable to initiate tumour growth to doses that always initiated tumour formation (Table 2). The tumour-forming capacity and phenotypic appearance were the same for primary xenografts and tumours passaged into secondary and tertiary recipients. The similar

Table 1 | Patient and tumour characteristics

Patient number	Age/sex	Tumour site	Tumour stage	Tumour differentiation	Xenograft formation	$CD133^+$ in tumour (%)	$CD133^+$ in normal (%)
P1	73/M	Right colon	IIIB	Moderate	Yes	14.0	2.1
P2	80/F	Liver	IV	Moderate	Yes	5.2	
P3	73/M	Liver	IV	Poor	Yes	14.7	
P4	70/F	Liver	IV	Moderate	Yes	1.8	
P5	74/F	Paracolic	IV	Poor	Yes	24.5	
P6	64/M	Liver	IV	Moderate	Yes	6.3	
P7	75/F	Right colon	I	Well	Yes	9.3	1.2
P8	34/F	Right colon	IIIC	Well	Yes	7.5	0.4
P9	63/F	Liver	IV	Well to moderate	Yes	12.1	
P10	66/F	Right colon	IIIC	Well to moderate	Yes	8.9	1.3
P11	74/F	Liver	IV	Moderate	Yes	19.0	
P12	65/F	Right colon	IIIC	Poor	Yes	15.9	0.85
P13	84/F	Sigmoid	I	Moderate	Yes	12.0	1.9
P14	58/M	Liver	IV	Moderate	Yes	17.6	
P15	53/M	Liver	IV	Moderate	Yes	18.2	
P16	75/F	Liver	IV	Not stated	Yes	10.4	
P17	56/M	Liver	IV	Moderate	Yes	3.2	

NOD/SCID mice were injected with bulk colon cancer cells from each tumour. All seventeen tumours generated xenografts in NOD/SCID mice. Ten of the tumours were carried through to tertiary passages in mice (tumours 7–15, and 17). Of the remaining tumours all except three (tumours 1, 2 and 16) were carried through to secondary mice. CD133 expression was determined by flow cytometry for each tumour before implantation. For the six primary colonic tumours, CD133 expression was also determined for normal colonic tissue.

[1]Division of Cell and Molecular Biology, University Health Network, Toronto, Ontario, M5G 1L7, Canada. [2]Department of Pathology and Laboratory Medicine, [3]Center for Cancer Genetics-Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto Ontario, M5G 1X5, Canada. [4]Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, M5S 1A8, Canada.

behaviour of the primary and passaged tumours made it possible to combine data to calculate the average frequency of CC-ICs in these tumours using the maximum-likelihood estimation method of limiting dilution assay[14,15]. We calculated that on average there was one CC-IC per 5.7×10^4 (95% confidence interval: one per 3.4×10^4 to one per 9.3×10^4) unfractionated colon cancer cells, although the limited number of recipients used for any single sample prevented precise estimation of the patient-to-patient variation. Thus, as has been shown for breast[16] and brain[17] cancer, only a small subset of colon cancer cells are able to initiate tumour growth.

By combining the quantitative assay with cell fractionation, we were able to test whether human colon cancer adheres to the stochastic model, in which every tumour cell has equal tumour initiation potential[6,7,18], or to the cancer stem cell (CSC) model, in which some cell fractions are enriched for CC-IC activity while others are completely devoid of CC-ICs[6,7,18] .We focused on fractionation based on CD133 expression. The phylogenetically conserved protein CD133 was recently identified[19,20] as a potential CSC marker in brain[17] and prostate[19] cancer. CD133 expression ranged from 1.8 to 24.5% in the colon cancer samples described in Table 1 (Fig. 2a). We used immunohistochemistry to show that CD133 was expressed in clusters amid negative cells (Fig. 2b). Normal colon tissue also expressed CD133 but at much lower levels than primary colonic tumours (0.4–2.1% normal versus 8.9–15.9%) (Table 1; Fig. 2a). To determine whether CD133 expression enriches for CC-ICs, colon cancer cells were separated into $CD133^-$ and $CD133^+$ fractions and injected into NOD/SCID mice. Of 47 mice (dose range: 2×10^3 to 2.5×10^5) injected with $CD133^-$ cells, only one mouse transplanted with the highest cell dose (Table 2) generated a tumour. Because the $CD133^-$ fraction was contaminated with 5–15% of cells expressing only low levels of CD133, we conclude that neither $CD133^-$ nor $CD133^{low}$ cells possess CC-IC activity. In contrast, tumours were consistently generated after injection of 1×10^3 colon cancer cells expressing the highest levels of CD133 ($CD133^+$), and injection of 100 $CD133^+$ cells resulted in tumour growth in one of four mice. Thus, while significantly enriched, not every $CD133^+$ cell represents

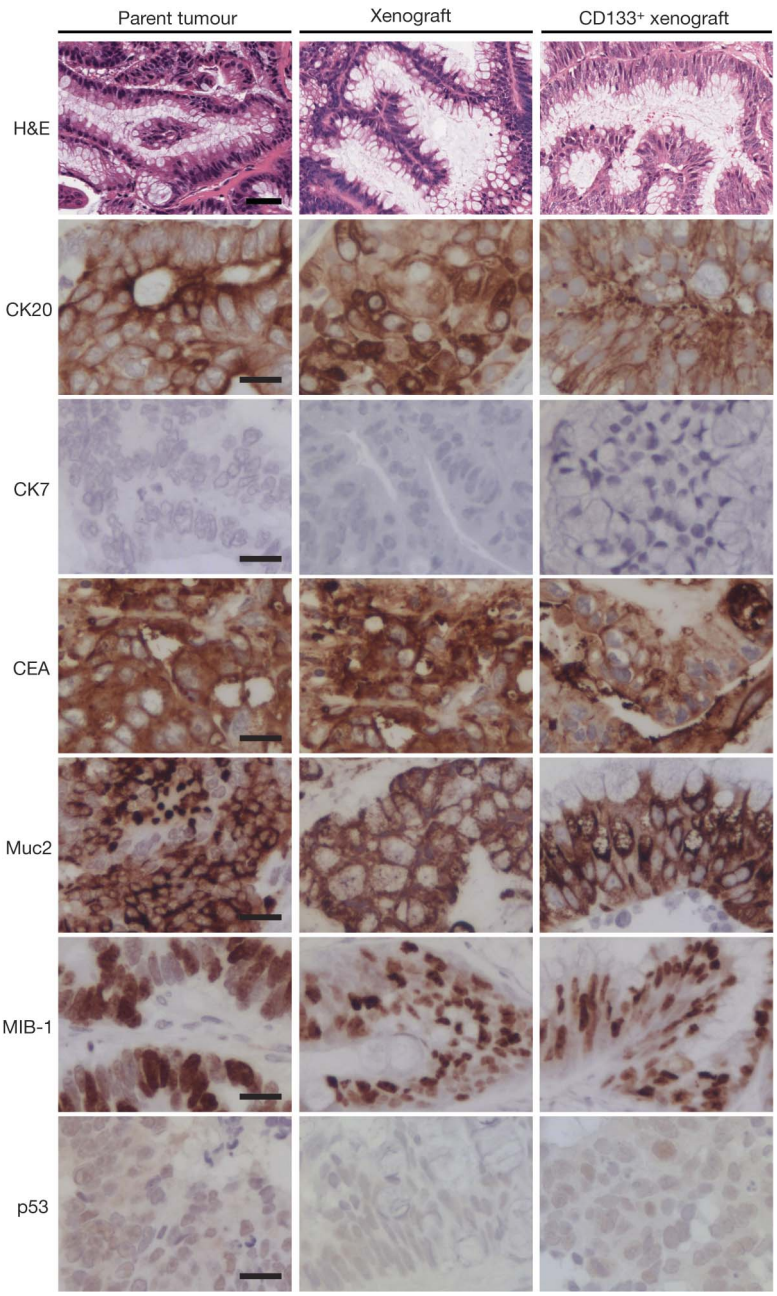


Figure 1 | Xenografts generated from both bulk and $CD133^+$ colon cancer cells resemble the original patient tumour. The parent tumour (tumour 14) is compared with xenografts generated from both primary and secondary passages of the tumour. The initial passage represents a xenograft generated from the injection of 1×10^5 bulk human colon cancer cells. The secondary xenograft was generated from the injection of 500 $CD133^+$ colon cancer cells. The histology of the three tumours, as expressed by H&E (haematoxylin and eosin) staining, shows well to moderately differentiated mucinous adenocarcinomas with intestinal differentiation including numerous goblet cells and intraluminal mucin. The immunohistochemical markers (including CK-20, CK7, CEA, Muc2, MIB-1 and p53) reveal comparable staining patterns in both the bulk and $CD133^+$ xenografts, as compared to the parent tumour. Images for each stain are taken at the same magnification. Scale bar represents 50 μm for H&E and 20 μm for all other stains.

a CC-IC. In total, 45 of 49 mice injected with $CD133^+$ cells developed tumours (Table 2). All tumours generated from $CD133^+$ cells were phenotypically similar to the original tumours (Fig. 1). Moreover, CD133 expression ranged from 1.7% to 22.4% in the xenografts, similar to the range seen in the original tumours. The isolation of tumorigenic and non-tumorigenic fractions, based on CD133 expression, provides strong support for the cellular organization of human colon cancer according to the CSC model.

Another prediction of the CSC model is that CC-ICs should self-renew to generate new CSCs and differentiate to generate non-tumorigenic progeny. Serial transplantation experiments from ten primary xenografts demonstrated that only $CD133^+$ and not $CD133^-$ cells were able to initiate tumour growth in serially transplanted secondary and tertiary mice. Tumours, either primary or passaged, could have been infiltrated with non-malignant cell types, but the high proportion (>98%) of $CD133^-$ cells that co-expressed the human-specific protein epithelial specific antigen from both primary and passaged tumours confirmed they were human colon cancer cells and not infiltrating murine cells or non-epithelial human cells that had somehow been co-passaged (Supplementary Fig. 2). Additionally, we showed that $CD133^-$ cells remained viable and stained positive for epithelial specific antigen under the renal capsule but were unable to regenerate tumours for as long as 15 to 21 weeks post-injection (Supplementary Fig. 3a, b and c). Furthermore, cells positive for epithelial specific antigen were also malignant, staining positive for p53, in cases where the parent tumours were $p53^+$ (Supplementary Fig. 3d). These studies were performed on passaged tumours, so it is highly unlikely the $CD133^-$ cells are pre-malignant cells, rather than malignant cells. Therefore, the $CD133^-$ cells are generated from the $CD133^+$ cells. Thus we can conclude that only $CD133^+$ CC-ICs can be serially passaged, forming xenografts that re-establish tumour heterogeneity, generating both $CD133^+$ and $CD133^-$ progeny in a ratio similar to that in the patient tumour (Fig. 2c).

To determine the frequency of CC-ICs within the $CD133^+$ subset we carried out a limiting dilution assay, using the same principles as described for unfractionated tumour cells[14,15]. The passaged xenografts matched the phenotype and tumour-forming capacity of the parent tumours, enabling us to combine data from passaged and primary cells. The frequency was calculated to average one CC-IC in 262 $CD133^+$ colon cancer cells (95% confidence interval: one in 129 to one in 534), representing a 216-fold enrichment of CC-ICs compared to unfractionated colon cancer cells.

Interestingly, the estimate of CC-IC frequency when back-calculated to take into account the proportion of $CD133^+$ cells within the unfractionated tumour is ~20-fold higher than when unfractionated cell suspensions were assayed. For example, multiplication of the CC-IC frequency (one in 262) by the mean level of CD133 expression for all samples (12%) yields an estimate of 20 CC-ICs per 57,000 unfractionated tumour cells, instead of the one in 57,000 measured in the initial limiting dilution assay. One possible explanation for this finding is that the $CD133^-$ progeny are negatively regulating the growth of the $CD133^+$ CC-IC fraction, thereby requiring a greater overall number of $CD133^+$ cells to give rise to a tumour, as has been observed in human haematopoietic stem cells[21].

Here we have identified and characterized CC-ICs from human colon tumour samples on the basis of their ability to initiate human colon cancer after transplantation into NOD/SCID mice. CC-ICs possessed two key criteria that define stem cells: after transplantation at limit dilution, single CC-ICs proliferated extensively and differentiated to produce tumours that were phenotypically similar to the original patient tumours, and as a population they self-renewed, enabling re-establishment of colon cancer in secondary and tertiary recipient mice. CC-ICs were almost exclusively $CD133^+$, while the $CD133^-$ fraction that comprised 81–98% of the tumour mass had no CC-IC activity. Thus colon cancer, like acute myelogenous leukaemia[22], breast[16] and brain[17] cancer is organized as a hierarchy in which a small population of CSCs sustain the tumour. The calculated frequency of CC-ICs was, on average, one in 262 $CD133^+$ cells, so clearly the majority of $CD133^+$ cells are not CC-ICs. As described for CD34 expression on acute myelogenous leukaemic stem cells, this result suggests there may be a hierarchy of CC-ICs and progenitors[23]. Thus, future studies using additional cell surface markers in combination with CD133 are necessary to purify the CC-IC fraction further. Finally, clonal tracking studies need to be carried out to establish self-renewal at the single cell level and determine whether different subclasses of CC-ICs exist[23].

Although we found $CD133^+$ CC-ICs in primary and metastatic tumours, most primary colon cancers tested were derived from right-sided tumours and may not be representative of all forms of colon cancer. Nevertheless, our findings should stimulate future studies directed towards increasing the range of colon cancer samples tested and addressing whether qualitative or quantitative CC-IC differences have prognostic value. Analysis of CC-ICs using molecular genetic techniques should further our understanding of the genetic abnormalities commonly associated with colon cancer, such as microsatellite status.

Table 2 | Limiting dilution analysis of the human colon cancer initiating cell

Colon cancer cell source	Cell dose	Number of samples tested	Identification numbers of samples tested	(Number of primary mice with tumours)/ (total number injected)	(Number of secondary mice with tumours)/ (total number injected)	Total number of mice with tumours (%)
Bulk	1×10^4	8	3–5,10–14	0/4	0/4	0/8 (0)
	2.5×10^4	8	4,6,10–14,17	1/6	0/2	1/8 (12.5)
	5×10^4	10	3–11,15	2/5	2/5	4/10 (40)
	7.5×10^4	8	3,5,7–9,11–13	4/8		4/8 (50)
	1×10^5	10	3–5,7–13	6/6	4/4	10/10 (100)
	1×10^6	17	1–17	17/17		17/17 (100)
	2×10^6	8	6–9,12–17	8/8		8/8 (100)
$CD133^+$	1×10^2	4	7,8,14,15		1/4	1/4 (25)
	5×10^2	6	5,6,11,13,14,17	1/1	4/5	5/6 (83.33)
	1×10^3	7	5–8,10,12,17	1/1	6/6	7/7 (100)
	5×10^3	8	8–13,15,17	1/1	7/7	8/8 (100)
	1×10^4	10	5,7–14,17	1/1	9/9	10/10 (100)
	2×10^4	9	5–7,9–11,13–15		9/9	9/9 (100)
$CD133^-$	5×10^3	5	6,9,11,12,17		0/5	0/5 (0)
	1×10^4	6	8,10,12,14,15,17	0/1	0/5	0/6 (0)
	2×10^4	6	5–7,9,10,13	0/1	0/6	0/6 (0)
	5×10^4	8	5,7–9,11,12,14,15	0/1	0/7	0/8 (0)
	1×10^5	8	6,8,10,12–15,17	0/1	0/7	0/8 (0)
	2.5×10^5	9	5,7–9,11,13–15,17		1/9	1/9 (11.1)

NOD/SCID mice were transplanted with: bulk ($n = 61$), $CD133^+$ ($n = 49$), and $CD133^-$ ($n = 47$) human colon cancer cells. All doses are displayed with the exception of 2×10^3 for $CD133^+$ ($n = 5$) and $CD133^-$ ($n = 5$), in which tumour-formation rates were 100% and 0%, respectively. Mice were killed at 6–21 weeks post-injection. Mice were considered negative if no tumour tissue was identified. Only doses that resulted in a mix of positive and negative mice were used to calculate the limiting dilution experiments.

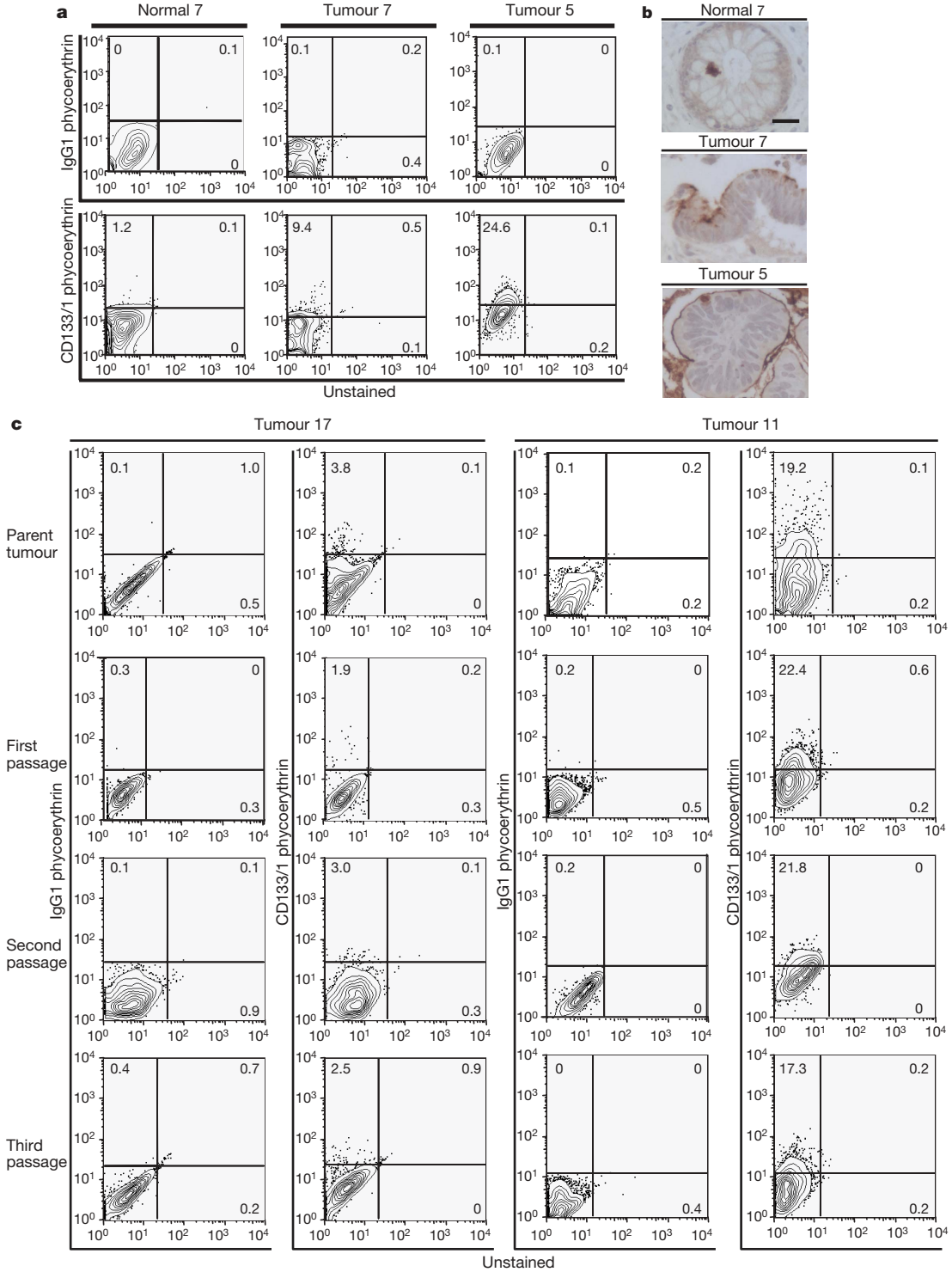


Figure 2 | Expression of CD133 in tumour and normal colonic tissue. a, Flow cytometric contour plots demonstrating the variable expression of CD133 between normal colon tissue (normal 7) and colon cancer tissue from the same patient (tumour 7) and a representative third tumour (tumour 5) showing higher CD133 expression. The upper and lower panels depict isotype controls and CD133 staining, respectively. **b**, Immunohistochemical staining for CD133: normal 7, and tumours 7 and 5 (all images are taken at the same magnification; scale bar represents 20 μm). **c**, Flow cytometric contour plots demonstrate preservation of CD133 expression through primary, secondary and tertiary passages as exemplified by tumours 17 and 11. Tumours from each passage were stained with CD133 phycoerythrin and an isotype-specific antibody (IgG1 phycoerythrin) and the proportion of $CD133^+$ cells is shown in each quadrant. CD133 expression varied between tumours 11 and 17. Each tumour maintained consistent levels of CD133 expression through three separate tumour passages. The first passage represented injection of bulk colon cancer cells but subsequent passages involved the isolation and injection of $CD133^+$ colon cancer cells.

Furthermore, as our understanding of normal colon stem and progenitor cell biology improves, it should be possible to gain insight into the cells that are the origin of colon cancer and the cellular context within which the well-characterized sequence of genetic events occurs[24,25].

The existence of tumorigenic and non-tumorigenic cells within colon cancers implies that not all the cells within a tumour are able to initiate and sustain neoplastic growth. This concept has important therapeutic implications, and may explain the observation that small numbers of disseminated cancer cells can be detected in the circulation of patients that never develop metastatic disease[18]. The identification of CC-ICs provides a powerful tool with which to develop a better understanding of tumour progression and the metastatic process, given that the CSC model predicts that the unit of selection in tumour progression would be the CSC itself. Moreover, because CC-ICs are the driving force sustaining tumour growth, developing adjuvant therapies directed at specifically eliminating the CC-IC fraction may prove to be a more effective strategy for reducing both local and distant recurrence[6,26]. The model described here will provide the means of further purifying and functionally characterizing the biological properties of the CC-IC fraction, with the goal of developing new therapeutic strategies directed specifically against CC-ICs.

METHODS

More detailed methods are in the Supplementary Information.

Tumour cell preparation. Colon cancer specimens were obtained from consenting patients, as approved by the Research Ethics Board at The University Health Network in Toronto. Tumour tissue was mechanically dissociated and incubated with Collagenase Type IV (Sigma) followed by magnetic bead separation to remove dead cells (Miltenyi Biotec).

Magnetic cell sorting and flow cytometry. Human colon cancer cells were magnetically labelled and separated by double passage using a CD133 Cell Isolation Kit (Miltenyi Biotec). Before separation, samples were assessed using a FACSCalibur flow cytometer (BD Biosciences), mouse IgG1s conjugated to phycoerythrin or fluorescein isothyiocyanate were used as isotype controls (BD Biosciences). CD133 expression was assessed using anti-CD133/1 phycoerythrin (Miltenyi-Biotec). To confirm the cells as human colon cancer, samples were tested using anti-epithelial specific antigen fluorescein isothyiocyanate (Biomeda) (Supplementary Fig. 2). At least 10,000 events were acquired for each sample and all cells positive for propidium iodide were gated out. After magnetic bead separation, samples were assessed by flow cytometry for purity.

Transplantation of human colon cancer cells into NOD/SCID mice. NOD/LtSz-scid/scid (NOD/SCID) mice were bred and maintained under defined conditions at the Ontario Cancer Institute under conditions approved by the Animal Care Committee of the Ontario Cancer Institute. Colon cancer cells were suspended in a 1:1 mixture of media and matrigel (BD Biosciences) and injected under the renal capsule of mice (8 weeks of age) that were sublethally irradiated (350 centiGray). Mice were anaesthetized while cells were injected under the renal capsule. All mice were killed when the tumour measured 1 cm, at the first sign of suffering, or between 15 and 21 weeks post-transplantation.

Received 21 August; accepted 26 October 2006.
Published online 19 November 2006.

1. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal cancer tumorigenesis. *Cell* **61,** 759–767 (1990).
2. Fearon, E. R. & Jones, P. A. Progressing toward a molecular description of colorectal cancer development. *FASEB J.* **6,** 2783–2790 (1992).
3. Radtke, F. & Clevers, H. Self-renewal and cancer of the gut: two sides of a coin. *Science* **307,** 1904–1909 (2005).
4. Nelson, H. *et al.* Guidelines 2000 for colon and rectal cancer surgery. *J. Natl Cancer Inst.* **93,** 583–596 (2001).
5. O'Connell, J. B., Maggard, M. A. & Ko, C. Y. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *J. Natl Cancer Inst.* **96,** 1420–1425 (2004).
6. Dick, J. E. Breast cancer stem cells revealed. *Proc. Natl Acad. Sci. USA* **100,** 3547–3549 (2003).
7. Wang, J. C. & Dick, J. E. Cancer stem cells: lessons from leukemia. *Trends Cell Biol.* **15,** 494–501 (2005).
8. Pocard, M., Tsukui, H., Salmon, R. J., Dutrillaux, B. & Poupon, M. F. Efficiency of orthotopic xenograft models for human colon cancers. *In Vivo* **10,** 463–469 (1996).
9. Ravi, R. *et al.* Elimination of hepatic metastases of colon cancer cells via p53-independent cross-talk between irinotecan and Apo2 ligand/TRAIL. *Cancer Res.* **64,** 9105–9114 (2004).
10. Golas, J. M. *et al.* SKI-606, a Src/Abl inhibitor with *in vivo* activity in colon tumor xenograft models. *Cancer Res.* **65,** 5358–5364 (2005).
11. Sack, M. J. & Roberts, S. A. Cytokeratins 20 and 7 in the differential diagnosis of metastatic carcinoma in cytologic specimens. *Diagn. Cytopathol.* **16,** 132–136 (1997).
12. Ishida, H. *et al.* Ki-67 and CEA expression as prognostic markers in Dukes' C colorectal cancer. *Cancer Lett.* **207,** 109–115 (2004).
13. Liang, J. T. *et al.* Microvessel density, cyclo-oxygenase 2 expression, K-ras mutation and p53 overexpression in colonic cancer. *Br. J. Surg.* **91,** 355–361 (2004).
14. Porter, E. H. & Berry, R. J. The efficient design of transplantable tumour assays. *Br. J. Cancer* **17,** 583–595 (1964).
15. Wang, J. C., Doedens, M. & Dick, J. E. Primitive human hematopoietic cells are enriched in cord blood compared with adult bone marrow or mobilized peripheral blood as measured by the quantitative *in vivo* SCID-repopulating cell assay. *Blood* **89,** 3919–3924 (1997).
16. Al Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100,** 3983–3988 (2003).
17. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432,** 396–401 (2004).
18. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
19. Marzesco, A. M. *et al.* Release of extracellular membrane particles carrying the stem cell marker prominin-1 (CD133) from neural progenitors and other epithelial cells. *J. Cell Sci.* **118,** 2849–2858 (2005).
20. Corbeil, D. *et al.* The human AC133 hematopoietic stem cell antigen is also expressed in epithelial cells and targeted to plasma membrane protrusions. *J. Biol. Chem.* **275,** 5512–5520 (2000).
21. Madlambayan, G. J. *et al.* Dynamic changes in cellular and microenvironmental composition can be controlled to elicit *in vitro* human hematopoietic stem cell expansion. *Exp. Hematol.* **33,** 1229–1239 (2005).
22. Lapidot, T. *et al.* A cell initiating human acute myeloid leukemia after transplantation into SCID mice. *Nature* **367,** 645–648 (1994).
23. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunol.* **5,** 738–743 (2004).
24. Al Hajj, M. & Clarke, M. F. Self-renewal and solid tumor stem cells. *Oncogene* **23,** 7274–7282 (2004).
25. Polyak, K. & Hahn, W. C. Roots and stems: stem cells in cancer. *Nature Med.* **12,** 296–300 (2006).
26. Al Hajj, M., Becker, M. W., Wicha, M., Weissman, I. & Clarke, M. F. Therapeutic implications of cancer stem cells. *Curr. Opin. Genet. Dev.* **14,** 43–47 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We gratefully acknowledge assistance from F. Meng, H. Begley and C. Ash for tissue acquisition, D. Hedley for advice on establishment of the xenograft model, J Wang for assistance with manuscript preparation, and the Dick laboratory members, P. Dirks and D.Hill for comments on the manuscript. We also acknowledge K. So and the University Health Network Pathology Research Program for tissue sectioning and immunohistochemistry. This work was supported by: a clinician-scientist award (C.A.O'B.), and grants (J.E.D.) from the Canadian Institute of Health Research, as well as grants to J.E.D. from Genome Canada through the Ontario Genomics Institute, the Ontario Cancer Research Network with funds from the Province of Ontario, the Leukemia and Lymphoma Society, the National Cancer Institute of Canada with funds from the Canadian Cancer Society and the Terry Fox Foundation, and a Canada Research Chair (J.E.D.).

Author Contributions C.A.O'B. planned the project, carried out experimental work, analysed data and prepared the manuscript. A.P. provided pathology analysis. S.G. provided clinical information and human tissues. J.E.D. planned the project, analysed data, and prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.E.D. (jdick@uhnres.utoronto.ca).

LETTERS

Identification and expansion of human colon-cancer-initiating cells

Lucia Ricci-Vitiani[1], Dario G. Lombardi[2], Emanuela Pilozzi[3], Mauro Biffoni[1], Matilde Todaro[4], Cesare Peschle[1] & Ruggero De Maria[1,2]

Colon carcinoma is the second most common cause of death from cancer[1]. The isolation and characterization of tumorigenic colon cancer cells may help to devise novel diagnostic and therapeutic procedures. Although there is increasing evidence that a rare population of undifferentiated cells is responsible for tumour formation and maintenance[2–4], this has not been explored for colorectal cancer. Here, we show that tumorigenic cells in colon cancer are included in the high-density CD133$^+$ population, which accounts for about 2.5% of the tumour cells. Subcutaneous injection of colon cancer CD133$^+$ cells readily reproduced the original tumour in immunodeficient mice, whereas CD133$^-$ cells did not form tumours. Such tumours were serially transplanted for several generations, in each of which we observed progressively faster tumour growth without significant phenotypic alterations. Unlike CD133$^-$ cells, CD133$^+$ colon cancer cells grew exponentially for more than one year *in vitro* as undifferentiated tumour spheres in serum-free medium, maintaining the ability to engraft and reproduce the same morphological and antigenic pattern of the original tumour. We conclude that colorectal cancer is created and propagated by a small number of undifferentiated tumorigenic CD133$^+$ cells, which should therefore be the target of future therapies.

Cancer neural stem cells can be identified and isolated through the CD133 marker[4], which is expressed by normal primitive cells of the neural, haematopoietic, epithelial and endothelial lineages[5–7]. To investigate whether there exists a CD133$^+$ cell population in colon cancer, we used flow cytometry to analyse the immunophenotype of colonic tumour cells shortly after tissue dissociation. The vast majority of the samples analysed showed the presence of rare cells ($2.5 \pm 1.4\%$) clearly positive for CD133 (Fig. 1a and Table 1). These cells did not express cytokeratin 20 (CK20) (Fig. 1b), an intermediate filament protein whose presence is essentially restricted to differentiated cells from gastric and intestinal epithelium and urothelium[8]. To determine the anatomical location of CD133$^+$ cells in colon cancer, we used immunohistochemistry to analyse a number of colon cancer sections from six different patients. All the samples analysed showed similar results, with the presence of rare CD133$^+$ cells in areas of high cellular density (Fig. 1c). CD133 expression in normal colon tissues was extremely infrequent (barely detectable upon extensive analysis of histological sections; data not shown) as compared with the tumour tissue.

The increased number of CD133$^+$ cells in cancer samples may result from their oncogenic transformation. To evaluate the tumorigenic potential of colon CD133$^+$ cells, we compared the ability of tumour-derived CD133$^+$ and CD133$^-$ cells to engraft and give rise to subcutaneous tumours in severe combined immunodeficient (SCID) mice. After surgical resection, colorectal cancer tissues were dissociated into single cells that were separated by immunomagnetic selection or flow cytometry on the basis of CD133 expression. This procedure resulted in a considerable enrichment of CD133$^+$ cells ($>80\%$) and an effective negative selection ($\geq 99.8\%$) of CD133$^-$ cells (Fig. 1d). The analysis of the CD133$^-$ population revealed that many (35–75%) of these cells displayed several features of cancers from the gastrointestinal tract, such as carcinoembryonic antigen (CEA)[9,10] expression and adenomatous polyposis coli (APC)[11] or p53 mutation[12], which were not present in normal colon epithelial cells (Supplementary Fig. 1). Before testing its *in vivo* oncogenic potential, we analysed the CD133$^+$ population for the presence of haematopoietic and endothelial progenitors. CD133$^+$ cells were all negative for the pan-haematopoietic marker CD45, over 97% were positive for the epithelial marker Ber-EP4, and less than 2% were putative endothelial progenitors and CD31$^+$ (Fig. 1e).

While 10^5 CD133$^-$ colon cancer cells did not induce tumour formation, the injection of 10^6 unseparated cells or 3,000 CD133$^+$ cells resuspended in matrigel generated visible tumours after 4–5 weeks from the transplant (Fig. 1f), indicating that colon-cancer-initiating cells are CD133$^+$. With the exception of Dukes stage A tumours, which reportedly are not tumorigenic in the current models of immunocompromised mice[13], we obtained engraftment with low numbers of CD133$^+$ cells isolated from tumours of all the other stages (Table 1). Despite the higher number of CD133$^+$ cells present in 10^6 unseparated cells, tumour formation after the injection of the total colon cancer population was slower and less efficient than that obtained with purified CD133$^+$ cells (Fig. 1f and Table 1), in line with results reported for breast cancer stem cells[14]. Moreover, haematoxylin-eosin staining and microscopic analysis indicated that CD133$^+$-derived tumour xenografts consistently reproduced the primary tumour at histological level, including specific signature features infrequently observed in colorectal cancer, such as areas of eosinophilic secretions scattered in the neoplastic tissue (Fig. 1g). As expected, purified CD133$^-$CEA$^+$ colon cancer cells were unable to transfer the tumour in immunocompromised mice (Table 1 and Supplementary Fig. 2). Thus, the tumorigenic population in colon cancer is restricted to CD133$^+$ cells, which are able to reproduce the original tumour in permissive recipients.

Normal and neoplastic stem cells from neural and epithelial organs can be expanded as sphere-like cellular aggregates[4,15–17] in serum-free medium containing EGF and FGF-2. We cultivated the colonic cells that we obtained after dissociation of cancer tissues using such a proliferative medium for undifferentiated cells. After 4 weeks of culture, we obtained colon spheres formed by aggregates of exponentially growing undifferentiated cells (Supplementary Fig. 3a) from 5

[1]Department of Hematology and Oncology, Istituto Superiore di Sanità, Viale Regina Elena 299, Rome 00161, Italy. [2]Mediterranean Institute of Oncology, Via Penninazzo 7, Viagrande 95029, Catania, Italy. [3]Department of Laboratory Medicine and Pathology, Sant'Andrea Hospital, University 'La Sapienza', Via di Grottarossa 1037, Rome 00189, Italy. [4]Department of Surgical and Oncological Sciences, University of Palermo, Via Liborio Giuffrè 5, Palermo 90127, Italy.

of 15 tumours (Supplementary Table 1). Within ten passages, the doubling time of colon spheres was approximately ten days, which became less than 7 after 30 passages (Supplementary Fig. 3b), probably as a consequence of selection. CD133$^-$ colon cancer cells invariably died in such serum-free conditions (Supplementary Table 1 and data not shown), but could grow for about two weeks in serum-containing medium before gradually declining in number (Supplementary Fig. 3b). This standard culture of adherent cells did not allow the persistence of the CD133$^+$ population, whose presence was essentially undetectable after 10 days of culture (Supplementary Fig. 3c, d). As reported[18], the vast majority of cells obtained in this condition expressed CK20 (Supplementary Fig. 3e).

In contrast, cells grown as colon spheres remained CD133$^+$ and expressed negligible amounts of CK20 (Supplementary Fig. 3c, e). The ability to grow exponentially and the absence of CK20 suggested that CD133$^+$ colon spheres were aggregates of primitive cancer cells. To investigate whether expanded CD133$^+$ cells in tumour spheres maintained the tumorigenic potential, we injected 50 or 500 spheres subcutaneously into SCID mice and monitored the formation of tumours weekly. Although 10^6 differentiated primary colon cancer cells were not tumorigenic, the tumour spheres engrafted and generated tumours, which grew rapidly and required the mouse to be killed (Supplementary Fig. 3f, g). We noticed that injection of the higher number of spheres resulted in a faster appearance of the tumours without altering the cancer growth rate, which was independent of the number of transplanted spheres (Supplementary Fig. 3g).

To determine the differentiation potential of these CD133$^+$ cells, tumour spheres were cultivated without EGF and FGF-2 in the presence of 5% serum. After one day of culture, floating undifferentiated cells attached to the plastic, gradually migrating from tumour spheres and differentiating into large and adherent cells (Fig. 2a). Upon differentiation, colon cancer cells expressed CK20 and acquired a morphology closely resembling the major colon cancer cell

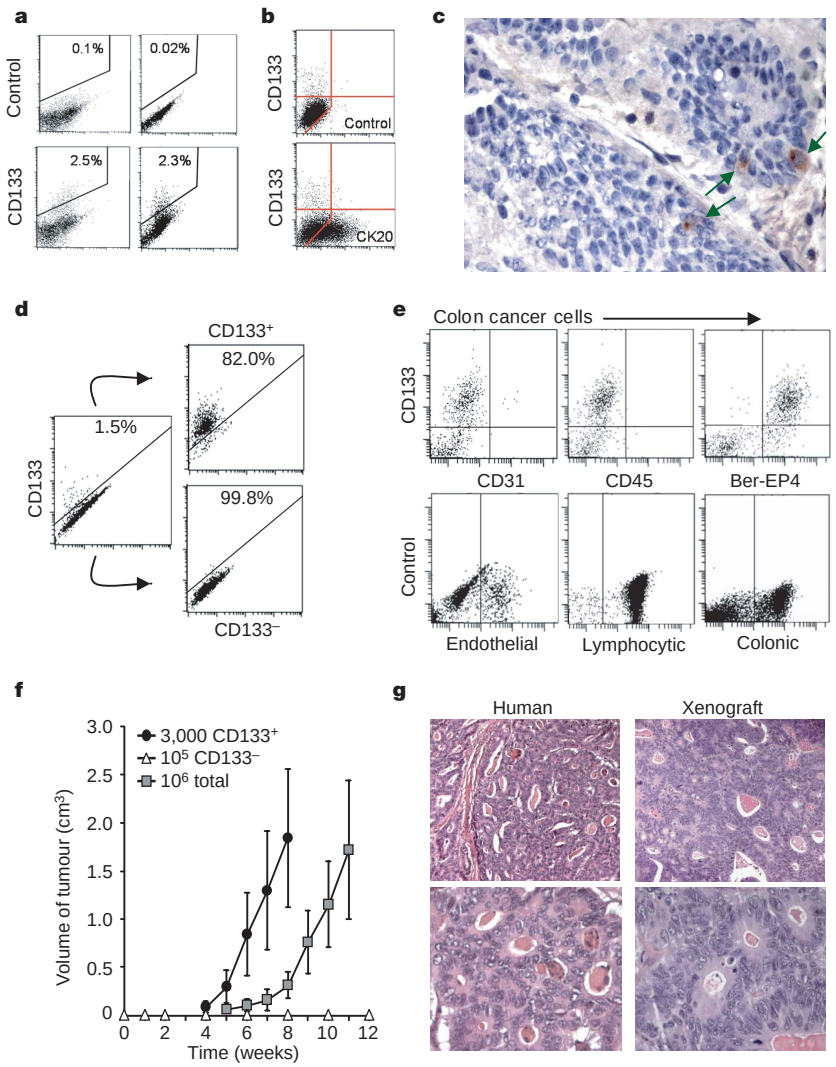


Figure 1 | A rare CD133$^+$ population of tumorigenic cells is present in colon cancer. **a**, **b**, Flow cytometry analysis of CD133 expression as single staining (**a**) or in combination with CK20 (**b**) in freshly dissociated colon adenocarcinoma cells derived from three representative tumours. **c**, Immunohistochemical analysis of CD133 expression in colon cancer. Positive cells are brown and indicated by arrows. **d**, CD133 expression before (left) and after sorting of positive (right, top) and negative (right, bottom) colon cancer cells. **e**, Flow cytometry analysis of CD133-enriched colon cancer cells. Freshly isolated normal endothelial, lymphocytic and colonic cells were used as positive controls. **f**, Evaluation of the tumorigenic potential of freshly isolated CD133$^+$, CD133$^-$ and unseparated (total) colon cancer cells after subcutaneous injection in matrigel. Data concerning tumour volumes generated by CD133$^+$ and CD133$^-$ cells are mean ± s.d. of five independent experiments in duplicate, referring to those tumours in which some of the 10^6 unseparated cells from at least one out of four injections were able to engraft. Data concerning unseparated cells are mean ± s.d. of 11 successful engraftments out of 40 injections. **g**, Haematoxylin-eosin analysis with different (100×, top; 400×, bottom) original magnifications of colon cancer sections from the original tumour (Human) and corresponding xenograft (Xenograft) obtained after injection of CD133$^+$ cells as in **e**.

population present in the original tumour (Fig. 2b). These differentiated cells expressed high levels of the caudal type homeobox transcription factor 2 (CDX2), a sensitive and specific marker for colorectal adenocarcinoma that was weakly expressed in the colon spheres (Fig. 2c–e)[19,20]. Comparative flow cytometry analysis of colon cancer cells from undifferentiated and differentiated tumour spheres showed that all cells expressed Ber-EP4 and CEA, whereas CD133 was significantly downregulated upon differentiation (Fig. 2f). During the differentiation process, $CD133^+$ cells lost their ability to transfer the tumour into immunocompromised mice (Fig. 2g and Supplementary Table 1), suggesting that colon-cancer-initiating cells need to remain undifferentiated to maintain tumorigenic potential. Thus, colon undifferentiated cancer cells can be cultured and expanded *in vitro* as colon spheres in proliferative serum-free medium containing growth factors. This property is common to neural and epithelial stem and progenitor cells, which grow as spherical aggregates that in the presence of serum or extracellular matrix differentiate upon growth factor removal[15–17].

This ability to obtain unlimited expansion of tumorigenic colon cancer cells could be exploited for more accurate preclinical studies provided that the expanded cells do not lose their ability to reproduce the original tumour. Although tumour cells may acquire genomic mutations after prolonged expansion, colon cancer spheres could be maintained in conditions of exponential growth for more than a year, without losing the ability to generate tumours. Tumour xenografts derived from colon spheres maintained in culture for one, six or twelve months closely reproduced the histological features of the original tumour, as indicated by haematoxylin-eosin staining and morphological analysis of tumour sections (Fig. 3a). Moreover, regardless of the duration of the *in vitro* expansion, tumours generated by colon spheres displayed a pattern of CDX2, β-catenin and CK20 identical to the primary tumour from which the cells were derived (Fig. 3b), indicating that expanded $CD133^+$ cells maintain their tumorigenic potential along with the ability to replicate the original tumour.

To investigate whether $CD133^+$ colon cancer cells display long-term tumorigenic potential, we evaluated the ability of these cells to generate tumours after serial transplantations. Tumour xenografts derived by the injection of freshly isolated $CD133^+$ cells were digested to isolate $CD133^+$ and $CD133^-$ cells, which in turn were

Table 1 | Case description and tumorigenic activity of $CD133^+$ colon cancer cells

Case	Age/Sex	Site	Grade	Dukes	CD133 expression	Number of cells injected			
						$CD133^+$	$CD133^-$	$CD133^-CEA^+$	Unseparated
1	68/M	Sigma	G3	C	2.4%	100,000 (2/2) 10,000 (1/2) 3,000 (2/2)	100,000 (0/2) 10,000 (0/2) 5,000 (0/2)		1,000,000 (2/4)
2	41/M	Left	G3	C	2.0%	10,000 (0/2) 5,000 (0/2) 3,000 (0/2)	100,000 (0/2) 10,000 (0/2) 5,000 (0/2)		1,000,000 (0/4)
3	69/F	Sigma	G2	D	2.5%	10,000 (2/2) 5,000 (2/2) 3,000 (2/2)	100,000 (0/2) 10,000 (0/2) 5,000 (0/2)		1,000,000 (3/4)
4	66/F	Left	G2	A	1.7%	10,000 (0/2) 5,000 (0/2) 3,000 (0/2)	100,000 (0/2) 10,000 (0/2) 5,000 (0/2)		1,000,000 (0/4)
5	67/M	Sigma	G2	C	1.4%	10,000 (2/2) 5,000 (1/2) 3,000 (1/2)	100,000 (0/2) 10,000 (0/2) 5,000 (0/2)		1,000,000 (0/4)
6	76/M	Right	G2	C	2.4%	10,000 (1/2) 5,000 (2/2) 3,000 (2/2)	100,000 (0/2) 10,000 (0/2)		1,000,000 (2/4)
7	68/M	Sigma	G2	C	1.8%	10,000 (2/2) 5,000 (2/2) 3,000 (2/2)	100,000 (0/2) 10,000 (0/2)		1,000,000 (3/4)
8	74/M	Sigma	G3	A	0.7%	5,000 (0/2) 3,000 (0/2) 1,500 (0/2)	100,000 (0/2) 10,000 (0/2)		1,000,000 (0/4)
9	58/M	Sigma	G2	A	2.4%	5,000 (0/2) 3,000 (0/2) 1,500 (0/2)	100,000 (0/2) 10,000 (0/2)		1,000,000 (0/4)
10	66/F	Right	G2	B	0.7%	5,000 (2/2) 3,000 (1/2) 1,500 (1/2)	100,000 (0/2) 10,000 (0/2)		1,000,000 (1/4)
11	79/F	Sigma	G2	A					
12	53/M	Right	G2	C					
13	70/F	Sigma	G2	B	2.6%	10,000 (1/2) 5,000 (2/2) 3,000 (1/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (1/4)
14	73/F	Right	G2	B	4.1%	10,000 (0/2) 5,000 (0/2) 3,000 (0/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (0/4)
15	68/M	Left	G2	C	2.7%	10,000 (2/2) 5,000 (2/2) 3,000 (2/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (2/4)
16	51/M	Sigma	G2	B	6.1%	10,000 (2/2) 5,000 (2/2) 3,000 (2/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (4/4)
17	53/M	Left	G2	A	1.7%	10,000 (0/2) 5,000 (0/2) 3,000 (0/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (0/4)
18	76/F	Sigma	G2	B	4.6%	10,000 (2/2) 5,000 (1/2) 3,000 (2/2)		100,000 (0/2) 10,000 (0/2)	1,000,000 (1/4)
19	63/F	Sigma	G2	D					

Freshly isolated subpopulations of colon cancer cells generated subcutaneous tumours in mouse xenografts. The success rate is shown in brackets.

transplanted into secondary mice. Although the CD133$^-$ population contained a majority of human colon cancer cells (Supplementary Fig. 4), only unseparated and CD133$^+$ cells were tumorigenic, whereas CD133$^-$ cells were not able to transfer the tumour into secondary mice (Fig. 3c); this confirms the data obtained with cells directly isolated from the human tumour. Moreover, CD133$^+$ tumour spheres obtained from similar CD133$^+$-derived primary xenografts were subsequently transplanted into secondary mice whose tumours were used as a new source of CD133$^+$ cells to generate tertiary and then quaternary tumours (Supplementary Table 1).

During the *in vivo* passages, CD133$^+$ cells did not lose their tumorigenic potential, but instead increased their aggressiveness, as indicated by the faster growth and increasing number of CD133$^+$CK20$^-$ cells of newly generated tumours (Fig. 3d, e). Accordingly, tumour spheres generated *in vitro* from these xenografts displayed an exponential growth that had the propensity to increase with the serial xenografting (Fig. 3f). Thus, the CD133$^+$ cell population resident in

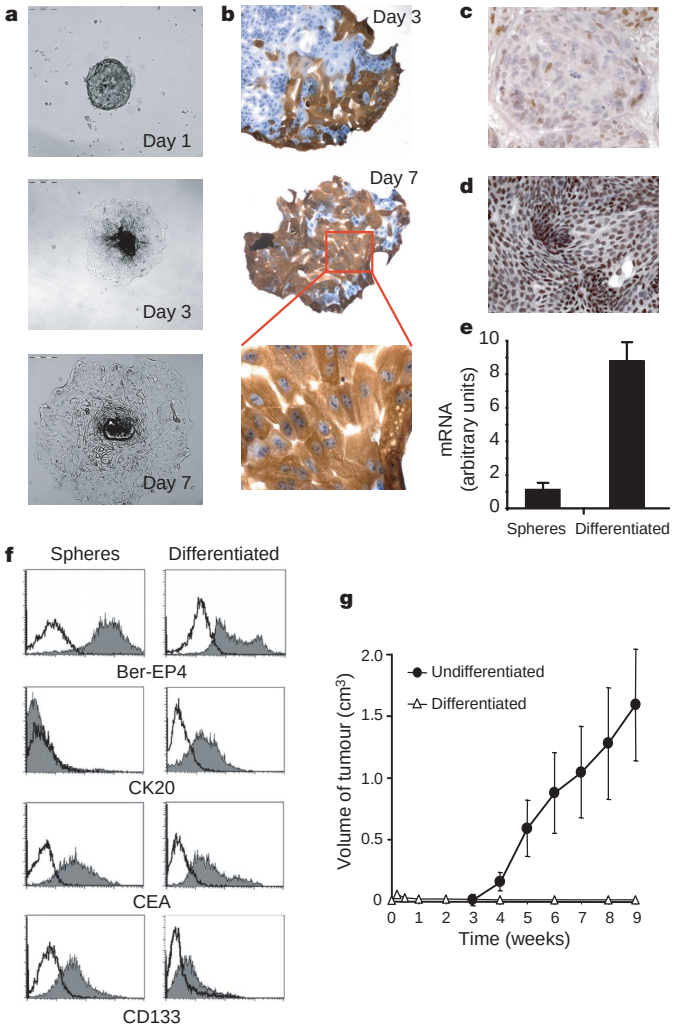


Figure 2 | The tumorigenic potential of CD133$^+$ colon cancer cells is lost upon differentiation. **a**, Microscopical analysis of colon cancer spheres cultivated in differentiation conditions for 1, 3 and 7 days. **b**, Immunocytochemical analysis of CK20 expression (brown) in differentiating colon cancer spheres. A higher magnification of the cells included in the red square is shown at the bottom. **c**, **d**, Immunocytochemical analysis of CDX2 expression (dark brown) in fibrin-included colon cancer spheres (**c**) and day 7 differentiated cells (**d**). **e**, Real-time polymerase chain reaction (PCR) analysis of CDX2 mRNA expression in colon cancer spheres and day 7 differentiated spheres. **f**, Flow cytometry analysis of colon cancer spheres and day 7 differentiated cells. Empty histograms represent isotype controls; grey histograms represent the specific binding for the indicated antigen. **g**, Tumorigenic potential of undifferentiated and differentiated CD133$^+$ colon cancer cells derived by spheres. Tumour volumes of mice injected with 5×10^5 cells are shown. Data are mean ± s.d. of three independent experiments in duplicate for patients (see cases in Table 1) 1, 3 and 7.

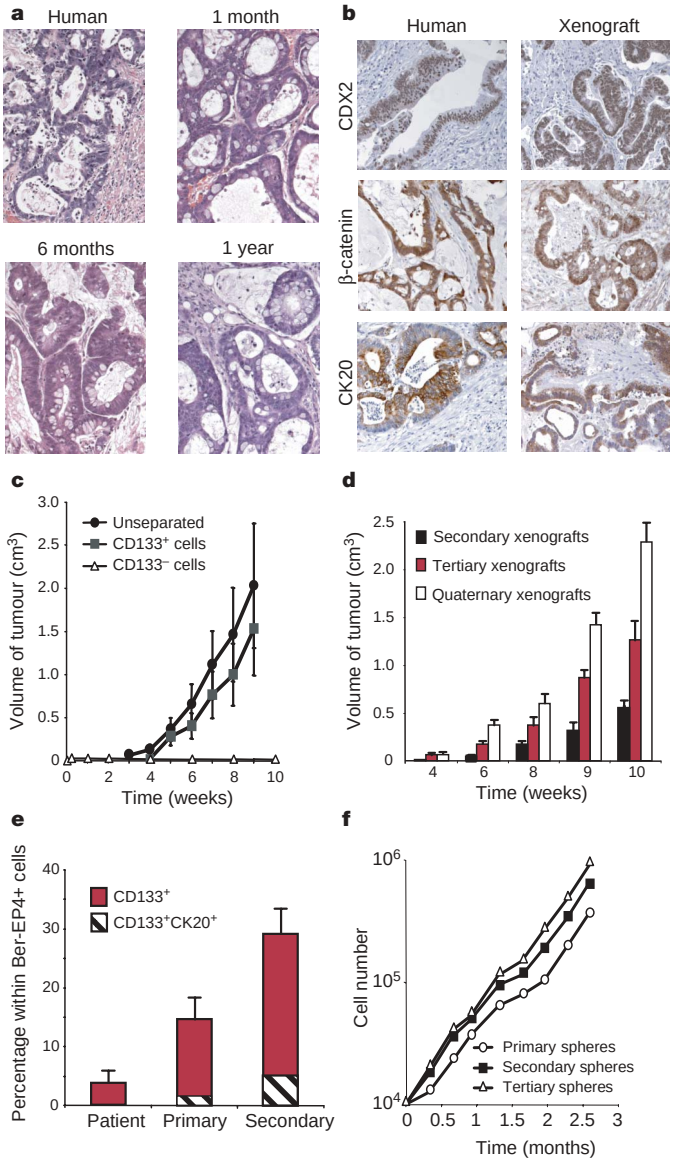


Figure 3 | Long-term tumorigenic potential of colon cancer CD133$^+$ cells. **a**, Haematoxylin-eosin analysis of original tumour (Human) and mouse xenografts generated by spheres expanded in culture for the indicated times. **b**, Immunohistochemical analysis of the original tumour (Human) and sphere-derived xenografts. **c**, Tumorigenic potential of 10^5 CD133$^+$, CD133$^-$ and total (unseparated) tumour cells purified from xenografts obtained in mice previously injected with freshly isolated CD133$^+$ cells. Data are mean ± s.d. of two independent experiments with patients (cases in Table 1) 15 and 16 in duplicate. **d**, Tumorigenic potential of colon cancer spheres derived by tumours induced by injection of 3,000 freshly isolated CD133$^+$ cells (Secondary xenografts), spheres derived by such secondary xenografts (Tertiary xenografts) and spheres derived by tertiary xenografts (Quaternary xenografts). Data are mean ± s.d. of three different experiments. **e**, Percentage of CD133$^+$ and CD133$^+$CK20$^+$ cells among BerEp4$^+$ cells from patient tumours and CD133$^+$-derived primary and secondary xenografts. Data are mean ± s.d. of three independent experiments. **f**, *In vitro* cell growth of colon cancer spheres obtained as indicated in **e**. A representative experiment of three is shown.

the colon tumour mass is able to generate serial xenografts showing a virtually unlimited growth potential.

Here, we demonstrated that tumorigenic colon cells are included in the rare undifferentiated population that expresses CD133. This antigen is a 120-kDa five-transmembrane-domain glycoprotein expressed on normal primitive haematopoietic, endothelial, neural and epithelial cells[5–7]. In adult and juvenile brain tumours, CD133 is a marker for cancer-initiating cells expressed by 6–29% of the total tumour population[4]. The cloning efficiency has not been determined for this tumorigenic colon cancer population, owing to the limitation of the *in vitro* system and the inability of single $CD133^+$ cells to grow in clonogenic assays. However, the small number of $CD133^+$ cells present in the colon cancer cell mass suggests that a significant proportion of these cells is tumorigenic and able to contribute to the disease progression.

Our data are in line with the cancer stem cell hypothesis that suggests that tumours are generated and maintained by a small subset of undifferentiated cells able to self-renew and differentiate into the bulk tumour population[21]. As in other cancer types, such as leukaemia[2], breast[14] and brain cancer[4], early progenitor or stem cells seem to be the target of oncogenic transformation in colon cancer. It is likely that these undifferentiated cells undergo symmetric and asymmetric divisions *in vivo*, resulting in the expansion of the tumorigenic cell population while producing a progeny of more differentiated cells that constitute the prevalent population of the tumour cell mass[22]. Thus, the molecular characterization of tumorigenic $CD133^+$ colon cancer cells is crucial for the development of new therapeutic strategies. Likewise, the possibility of obtaining a virtually unlimited expansion of colon cancer tumorigenic cells has considerable therapeutic implications for *in vitro* and *in vivo* evaluation of drug efficacy. In this context, the use of xenografts carrying a neoplastic lesion that closely resembles the original tumour seems more reliable than cell-line-based xenografts, and might be applied in the future for optimizing individualized therapies.

METHODS

Cell culture. Tumour samples were subjected to mechanical and enzymatic dissociation. The resulting cancer cells were cultured in a serum-free medium supplemented with 20 ng ml^{-1} EGF and 10 ng ml^{-1} FGF-2. To obtain primary tumour cell cultures, after enzymatic dissociation cells were plated onto collagen-coated dishes in DMEM medium containing 10% FCS. Endothelial cells were obtained by mechanical and enzymatic dissociation from a fragment of the human inferior thyroid vein.

Magnetic and cytofluorimetric cell separation. For magnetic separation, cells were labelled 24–48 h after enzymatic dissociation with CD133/1 microbeads using the Miltenyi Biotec CD133 cell isolation kit. Alternatively, cells were labelled with CD133/1-phycoerythrin antibody (Miltenyi Biotec) and sorted with a FACS Aria (Becton Dickinson). After magnetic or cytofluorimetric sorting, cell purity was evaluated by flow cytometry using CD133/2 (293C3)-phycoerythrin or CD133/2 (293C3)-APC antibodies (Miltenyi Biotec).

Transplantation of cancer cells. Unseparated, $CD133^+$ and $CD133^-$ purified populations were injected subcutaneously into the flanks of SCID mice. After 8–10 weeks mice were killed by cervical dislocation, tumours were removed, fixed in 10% neutral buffered formalin solution (Sigma) and paraffin embedded.

Immunohistochemistry. Immunohistochemistry was performed on formalin-fixed paraffin-embedded tissue, cell blocks or frozen tissue. Paraffin sections were dewaxed in xylene and rehydrated with distilled water. The slides were subsequently incubated with the following antibodies: CDX2 (BioGenex), CK20 (Dako), β-catenin (BD Transduction Laboratories). Cryostat sections were acetone-fixed and incubated at room temperature with anti-human CD133/1 (Miltenyi Biotec). The reaction was performed using Elite Vector Stain ABC systems (Vector Laboratories) and DAB substrate chromogen (DakoCytomation) followed by haematoxylin counterstaining.

Real-time PCR. The relative quantification of CDX2 messenger RNA was performed by *Taq*Man technology, using the ABI PRISM 7900 DNA sequence detection system and ready-to-use primers/probe mixes (Applied Biosystems). Human GAPDH was used as the housekeeping gene for the amplifications.

Received 20 September; accepted 30 October 2006.
Published online 19 November 2006.

1. Jemal, A. *et al.* Cancer statistics, 2006. *CA Cancer J. Clinic.* **56,** 106–130 (2006).
2. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3,** 730–737 (1997).
3. Pardal, R., Clarke, M. F. & Morrison, S. J. Applying the principles of stem-cell biology to cancer. *Nature Rev. Cancer* **3,** 895–902 (2003).
4. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432,** 396–401 (2004).
5. Uchida, N. *et al.* Direct isolation of human central nervous system stem cells. *Proc. Natl Acad. Sci. USA* **97,** 14720–14725 (2000).
6. Yin, A. H. *et al.* AC133, a novel marker for human hematopoietic stem and progenitor cells. *Blood* **90,** 5002–5012 (1997).
7. Salven, P., Mustjoki, S., Alitalo, R., Alitalo, K. & Rafii, S. VEGFR-3 and CD133 identify a population of CD34+ lymphatic/vascular endothelial precursor cells. *Blood* **101,** 168–172 (2003).
8. Moll, R. Cytokeratins as markers of differentiation in the diagnosis of epithelial tumors. *Subcell. Biochem.* **31,** 205–262 (1998).
9. Davidson, B. *et al.* Detection of malignant epithelial cells in effusions using flow cytometric immunophenotyping: an analysis of 92 cases. *Am. J. Clin. Pathol.* **118,** 85–92 (2002).
10. Sheahan, K. *et al.* Differential reactivities of carcinoembryonic antigen (CEA) and CEA-related monoclonal and polyclonal antibodies in common epithelial malignancies. *Am. J. Clin. Pathol.* **94,** 157–164 (1990).
11. Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359,** 235–237 (1992).
12. Rodrigues, N. R. *et al.* p53 mutations in colorectal cancer. *Proc. Natl Acad. Sci. USA* **87,** 7555–7559 (1990).
13. Jessup, J. M. *et al.* Growth potential of human colorectal carcinomas in nude mice: association with the preoperative serum concentration of carcinoembryonic antigen in patients. *Cancer Res.* **48,** 1689–1692 (1988).
14. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100,** 3983–3988 (2003).
15. Vescovi, A. L. *et al.* Isolation and cloning of multipotential stem cells from the embryonic human CNS and establishment of transplantable human neural stem cell lines by epigenetic stimulation. *Exp. Neurol.* **156,** 71–83 (1999).
16. Dontu, G. *et al. In vitro* propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* **17,** 1253–1270 (2003).
17. Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63,** 5821–5828 (2003).
18. Chu, P., Wu, E. & Weiss, L. M. Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms: a survey of 435 cases. *Mod. Pathol.* **13,** 962–972 (2000).
19. Ee, H. C., Erler, T., Bhathal, P. S., Young, G. P. & James, R. J. Cdx-2 homeodomain protein expression in human and rat colorectal adenoma and carcinoma. *Am. J. Pathol.* **147,** 586–592 (1995).
20. Witek, M. E. *et al.* The putative tumor suppressor Cdx2 is overexpressed by human colorectal adenocarcinomas. *Clin. Cancer Res.* **11,** 8549–8556 (2005).
21. Wang, J. C. & Dick, J. E. Cancer stem cells: lessons from leukemia. *Trends Cell Biol.* **15,** 494–501 (2005).
22. Clevers, H. Stem cells, asymmetric division and cancer. *Nature Genet.* **37,** 1027–1028 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Zeuner for discussions. This work was supported by grants from Associazione Italiana per la Ricerca sul Cancro and the Italian Health Ministry to R.D.M.

Author Contributions Experimental work and data analysis were done by L.R.-V., D.G.L., E.P., M.B. and M.T.; project planning and supervision was done by R.D.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D.M. (rdemaria@tin.it).

LETTERS

A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family

Nicolas Bocquet[1]*, Lia Prado de Carvalho[1]*, Jean Cartaud[2], Jacques Neyton[3], Chantal Le Poupon[1], Antoine Taly[1], Thomas Grutter[1], Jean-Pierre Changeux[1] & Pierre-Jean Corringer[1]

Ligand-gated ion channels (LGICs) mediate excitatory and inhibitory transmission in the nervous system. Among them, the pentameric or 'Cys-loop' receptors (pLGICs) compose a family that until recently was found in only eukaryotes. Yet a recent genome search identified putative homologues of these proteins in several bacterial species[1]. Here we report the cloning, expression and functional identification of one of these putative homologues from the cyanobacterium *Gloeobacter violaceus*. It was expressed as a homo-oligomer in HEK 293 cells and *Xenopus* oocytes, generating a transmembrane cationic channel that is opened by extracellular protons and shows slow kinetics of activation, no desensitization and a single channel conductance of 8 pS. Electron microscopy and cross-linking experiments of the protein fused to the maltose-binding protein and expressed in *Escherichia coli* are consistent with a homo-pentameric organization. Sequence comparison shows that it possesses a compact structure, with the absence of the amino-terminal helix, the canonical disulphide bridge and the large cytoplasmic domain found in eukaryotic pLGICs. Therefore it embodies a minimal structure required for signal transduction. These data establish the prokaryotic origin of the family. Because *Gloeobacter violaceus* carries out photosynthesis and proton transport at the cytoplasmic membrane[2], this new proton-gated ion channel might contribute to adaptation to pH change.

Iterative sequence profile searches in bacterial genomes[1] revealed fifteen putative homologues of eukaryotic pLGICs, which encompass an all-β-sheet hydrophilic domain and four transmembrane segments (M1–4; Fig. 1a). Because the bacterial sequences showed weak or no conservation with those amino acids in eukaryotic pLGICs that contribute to neurotransmitter binding (loops A–E, Fig. 1a) or to the ion channel, their ability to form ion channels gated by ligand was until now strictly hypothetical.

To address this issue, we cloned the gene from *Gloeobacter violaceus* (named *Glvi*): the closest homologue to eukaryotic pLGICs, sharing 20% amino acid identity with the human α7 nicotinic acetylcholine receptor (nAChR) subunit (Fig. 1a). We flanked the predicted mature Glvi protein-coding sequence with the signal peptide from the α7nAChR and a carboxy-terminal haemagglutinin (HA) epitope. Anti-HA immunofluorescence of human embryonic kidney 293 (HEK) cells transiently transfected with this Glvi complementary DNA showed a robust labelling at the cell surface (Fig. 2a), indicating an efficient expression and export of the protein, with its C terminus accessible from the extracellular compartment.

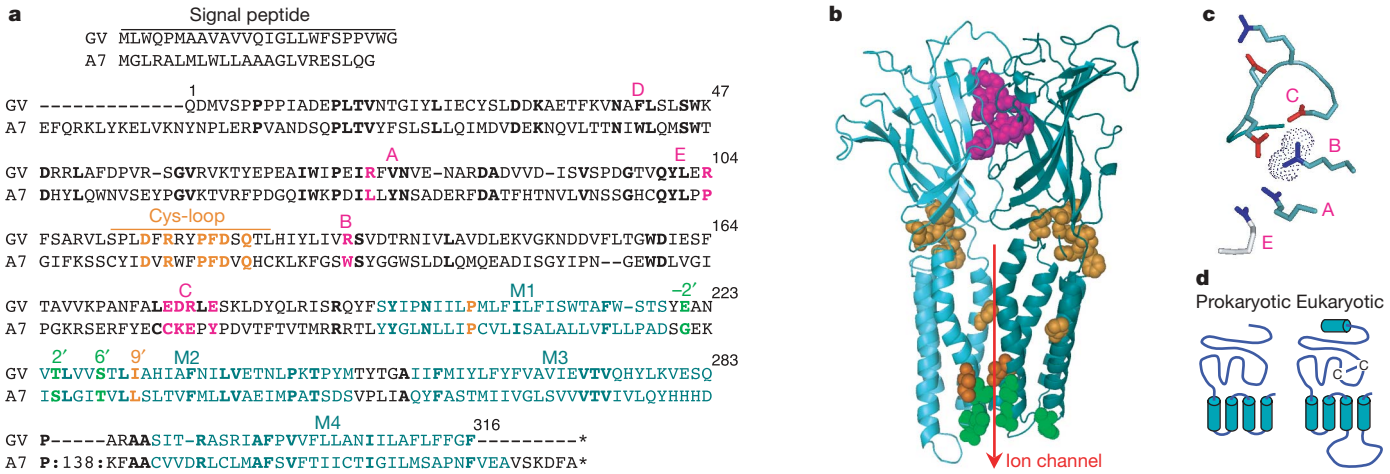


Figure 1 | The Glvi protein is related to eukaryotic pLGICs. **a**, Sequence alignment of the α7nAChR (A7) and Glvi (GV). Some α7 residues contributing to gating (orange) to the channel (green), and close to the agonist site (magenta) are coloured. *x*′ numbering takes as origin the predicted start of the M2 transmembrane segment. **b**, Homology model of the Glvi protein. Only two subunits are shown, and are oriented behind the ion pore represented as a red arrow. Key residues are represented as spheres, with the same colour code as in panel **a**. **c**, Closer view of Glvi charged residues that align with α7 residues close to the agonist-binding site. Carboxylates are represented in red and guanidino moieties in blue. **d**, Schematic topology of the Glvi protein compared with eukaryotic α7 nAChR subunit (cylinders, M1–M4).

[1]Unit of Receptor and Cognition, CNRS URA D2182, Pasteur Institute, 75015 Paris, France. [2]Cellular Biology of Membranes, CNRS UMR 7592, Institut Jacques Monod, Université Paris VI et VII, 75005 Paris, France. [3]Laboratory of Neurobiology, CNRS UMR 8544, Ecole Normale Supérieure, 75005 Paris, France.
*These authors contributed equally to this work.

It has been predicted that some bacterial homologues could be involved in chemotaxis[1]. Yet none of the compounds of a cocktail of sugars (*N*-Acetyl-glucosamine, L-arabinose, D-fructose, D-galactose, β-glycerophosphate, β-D-lactose, D-mannitol, mannose, sucrose) nor any of 13 representative amino acids, each applied at concentrations above 2 mM, elicited any significant current on whole-cell patch-clamp recording of Glvi-transfected HEK cells. In contrast, a proton concentration jump from pH 7.4 to pH 5.0 elicited robust currents two days after transfection (Fig. 3a). The onset of the response was slow and correctly fitted with a mono-exponential ($\tau = 260 \pm 140$ ms; all errors are s.d., $n = 14$), and its offset on return to pH 7.4 was rapid. The response to protons did not desensitize, even during a prolonged 30 s period at pH 5.0, generating a stable plateau (means inward, 5.7 ± 2.0 nA at -60 mV, $n = 14$; outward, 2.2 ± 0.8 nA at $+40$ mV, $n = 8$). Curves of pH response measured at the current plateau yielded a pH_{50} (the pH at which half maximal current is achieved) of 5.1 ± 0.1 ($n = 4$) with positive cooperativity ($n_H = 1.9 \pm 0.4$) (Fig. 3c, d).

Because HEK cells express endogenous acid sensing ion channels[3], we tested mock-transfected cells. The pH 5.0 solution elicited small currents that rapidly desensitized (at -60 mV, peak at 0.55 ± 0.65 nA and plateau at 0.11 ± 0.14 nA, $n = 12$; at $+40$ mV, plateau at 0.15 ± 0.18 nA, $n = 11$) (Fig. 3a). Thus, at the plateau, currents were 50- to 15-fold smaller than those recorded in Glvi-transfected cells, indicating a minor contribution to currents recorded in Glvi-transfected cells. Furthermore, the Glvi protein also generates proton-elicited currents in *Xenopus* oocytes, whereas non-injected oocytes showed either no response to pH 5.0 ($n = 5$), or small (<0.15 μA) and fluctuating currents that started around 5 s after the onset of pH 5 application ($n = 6$). Indeed, one day after Glvi plasmid injection, pH 5 elicited currents of 9.0 ± 2.5 μA at -60 mV ($n = 12$; Fig. 3b) that were similar in shape (slow onset, faster offset, no desensitization) and sensitivity ($pH_{50} = 4.9 \pm 0.5$, $n_H = 1.5 \pm 0.2$; Fig. 3d) to those recorded in transfected HEK cells. In addition, in Glvi-expressing oocytes, we found that tetraethylammonium (TEA) inhibited the proton-activated currents in a dose- and voltage-dependent manner (Fig. 3b, e)—two features expected for a channel blocker—whereas the blocker of acid sensing ion channels, amiloride, had no effect. In HEK cells expressing Glvi, TEA and QX-222 showed similar blocking properties, 1 and 10 mM QX-222 producing respectively $35 \pm 11\%$ and $83 \pm 5\%$ inhibition at pH 5 and -60 mV.

The ion selectivity of the Glvi channel was studied in transfected HEK cells. With NaCl and KCl external solutions, the reversal potential (E_{rev}) was close to zero (2.6 ± 2.9 mV ($n = 6$) and 0.8 ± 2.2 mV ($n = 5$), respectively). Replacing external NaCl by mannitol shifted E_{rev} to -56 ± 3 mV ($n = 6$), indicating a cationic channel with similar permeabilities for Na^+ and K^+ (Fig. 3f). This was confirmed by the replacement of external Cl^- by isethionate, which barely modified the E_{rev} (-4.5 ± 3.6 mV, $n = 5$) but caused a decrease of the currents, indicating a weak voltage-independent inhibitory effect.

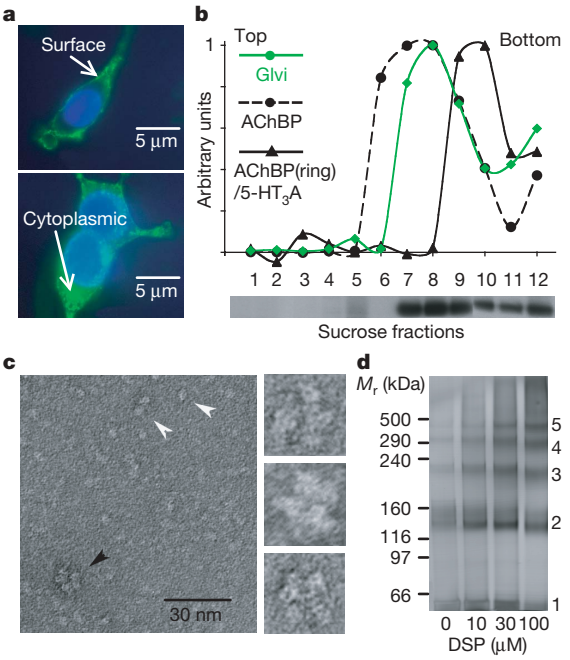


Figure 2 | Expression and oligomerization of the Glvi protein. **a**, Immunofluorescence imaging of HEK cells expressing the Glvi protein (green), in the absence (upper panel, intact cells) or presence (lower panel, permeabilized cells) of Triton X-100. Cell nuclei are coloured blue with DAPI. **b**, Typical 3–30% sucrose gradients. The metabolically labelled Glvi protein expressed in HEK cells was quantified in each fraction by autoradiography (SDS–PAGE below). AChBP(ring)/5-HT_3A and AChBP were quantified by [^{3}H]epibatidine binding and normalized. **c**, Electron microscopy of the negatively stained DDM-solubilized MBP-Glvi protein. White arrows indicate dissociated proteins, and the black arrow indicates proteins associated in a rosette. The inset shows typical rosettes at higher magnification (given by squares of 20×20 nm). **d**, Non-reducing 3–8% SDS–PAGE of the MBP-Glvi protein cross-linked with increasing concentrations of DSP. Putative mono- to pentamer cross-linked products are indicated.

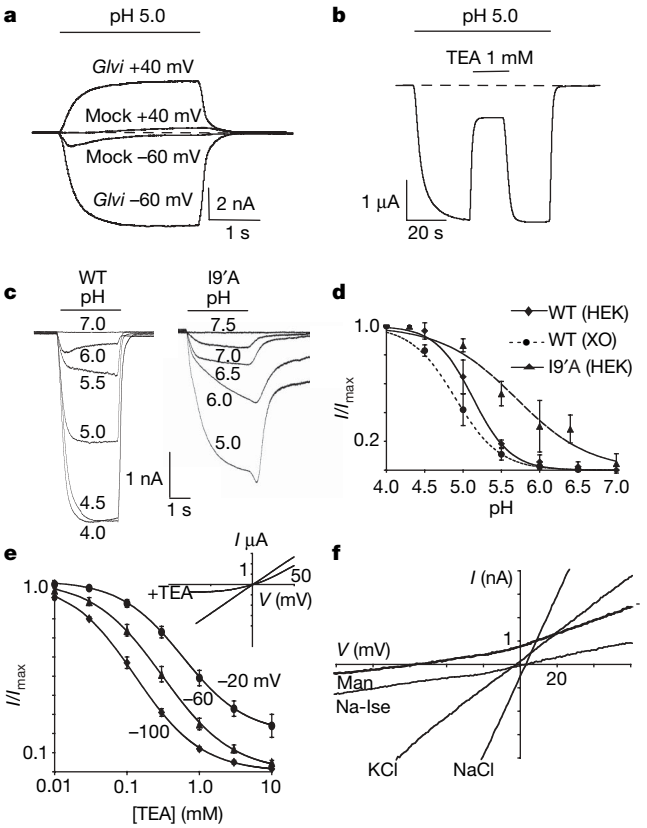


Figure 3 | The Glvi protein functions as a proton-gated ion channel. **a**, Whole-cell currents evoked by fast application of a pH 5.0 solution in HEK cells expressing the Glvi protein or GFP only (mock transfection). **b**, Two-electrode voltage-clamp currents evoked by superfusion of a pH 5.0 solution of a *Xenopus* oocyte injected with the Glvi plasmid, at -60 mV holding potential. **c**, Current traces evoked at different pHs on HEK cells transfected with Glvi (WT) or its mutant I9′A. For I9′A, pH applications were separated by 6 min rinsing intervals to ensure return of the current to baseline. **d**, pH-response curves at the current plateau in HEK cells and oocytes (XO) at -60 mV (mean of 3 experiments). **e**, TEA dose-inhibition curves of pH 5.0-evoked currents in oocytes expressing the Glvi protein at different holding potentials (mean of 3 experiments). Insert, *I–V* curves recorded in the presence (1 mM) and absence of TEA. **f**, *I–V* curves recorded at pH 5 in different extracellular solutions, on a single HEK cell expressing the Glvi protein. Error bars are s.d.

According to the Goldmann–Hodgkin–Katz equation, the dependence of E_{rev} as a function of the external NaCl concentration gave relative permeabilities $P_{Na}/P_K = 1.3 \pm 0.2$ and $P_{Cl}/P_K = 0$ (see Supplementary Information). The E_{rev} in external NaCl was independent of the applied pH (E_{rev} of 6.1 ± 1.0, 6.4 ± 0.6, 6.4 ± 0.6 and 8.4 ± 1.5 mV, at pH 4.0, 4.5, 5.0 and 5.5, respectively; $n=3$; see Supplementary Information), indicating a negligible contribution of proton flux to the current in this condition. Last, electrophysiological properties were unchanged on removal of the added C-terminal HA peptide.

In large outside-out patches of Glvi-transfected HEK cells, pH-evoked currents showed properties identical to those observed in the whole cell (similar pH_{50}, reversal potentials and blockade by TEA), whereas no currents were observed on mock-transfected cells (data not shown). In smaller patches (Fig. 4a), single-channel events were recorded at pH 5.0. Single-channel activity was also observed in cell-attached patches (Fig. 4b), with a unitary conductance (γ) of 8 ± 2 pS (Fig. 4c).

These data establish that Glvi functions as an LGIC, providing evidence, together with sequence homology, that it belongs to the pLGIC superfamily. Several experiments further support a homopentameric organization: (1) in sucrose gradients, the detergent-solubilized Glvi protein expressed in HEK cells sedimented as a symmetrical peak, followed by minor species of higher sedimentation velocity coefficients that probably corresponded to aggregated complexes (Fig. 2b). The Glvi peak (37 kDa per subunit, calculated from amino acid sequence) was closer to the pentameric acetylcholine-binding protein (AChBP) peak (35 kDa per subunit[4]) than to the pentameric chimera named AChBP (ring)/5-HT_3 protein (57 kDa per subunit, assuming identical glycosylations to that of AChBP[5]). (2) We examined by electron microscopy the Glvi protein fused to the maltose-binding protein from *E. coli* (MBP; 40.7 kDa), a construct that was efficiently expressed in *E. coli* (300 μg $litre^{-1}$) in contrast to Glvi, which resulted in arrest of cell growth. It revealed both dissociated proteins and proteins symmetrically associated in 'rosettes' that resemble *Torpedo marmorata* nAChRs[6], but with a diameter of approximately 12 nm (instead of 8 nm) (Fig. 2c). This difference in size is probably due to the fusion with MBP. (3) Cross-linking experiments with increasing concentrations of dithiobis (succinimidyl propionate) (DSP) resulted in the progressive appearance of four bands on an SDS–polyacrylamide gel, with molecular weights compatible with di-, tri-, tetra- and pentameric cross-linked products of MBP–Glvi (apparent molecular weight 63 kDa, calculated 77.6 kDa; Fig. 2d).

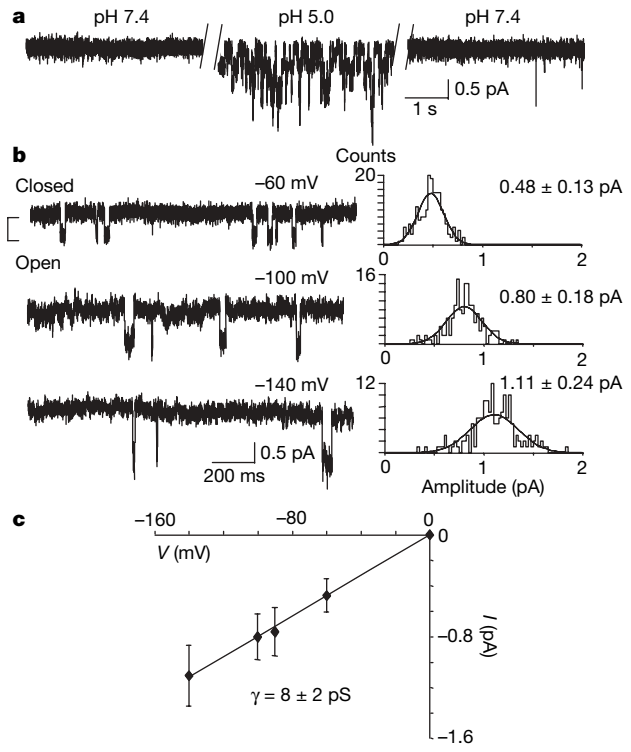


Figure 4 | Single-channel currents through Glvi protein in HEK cells. **a**, Outside-out patch at −60 mV displayed at 0.4 kHz bandwidth. **b**, Typical cell-attached currents activated at pH 6 and corresponding amplitude histograms from the same cell. Acquired at 10 kHz and displayed at 1 kHz. **c**, Current–voltage relationship for the patch shown in **b**. Error bars are s.d.

We built an illustrative 3D model of Glvi by homology using an α7nAChR model as a template[7], which was derived from the AChBP structure[8] and electron microscopy data[9,10] (Fig. 1b). In the channel domain lined by M2, some residues contributing to cation permeation are present in Glvi: Ser/Thr at position 2′ and 6′ (ref. 11), and −2′ Glu adjacent to −1′α7Glu. Glvi weakly discriminates between monovalent cations, but shows high charge-selectivity, properties that are typical of cationic pLGICs. This suggests a common organization of the channel between Glvi and eukaryotic pLGICs[12,13]. In addition, some key regions contributing to the gating process are conserved: (1) the 'Cys-loop', a DxRxxPFDxQ motif, flanked in eukaryotes by two canonical bridged cysteines that are missing in Glvi[14–16]; (2) the proline within M1 (GlviPro203; ref. 17); and (3) the equatorial ring of hydrophobic residues within M2 (Glvi Ile9'), in which mutation into Ala or Ser causes, for most pLGICs, an increased sensitivity to agonists, a weaker desensitization and more spontaneously active channels[18–20]. The homologous mutation Glvi Ile9′Ala had a striking phenotype: (1) the mutated protein was toxic for HEK cells which died 2 days after transfection; (2) low pH solutions elicited robust currents one day after transfection, but the offset of the response after activation was dramatically slowed down (5 s after the offset of pH 5.0 application, the current dropped to $22 \pm 15\%$ ($n = 10$) of the peak response, Fig. 3c); and (3) the pH_{50} was shifted to lower concentrations ($pH_{50} = 5.7 \pm 0.2$, $n_H = 1.2 \pm 0.4$ ($n = 4$); Fig. 3d). The increased sensitivity to the agonist and the slower deactivation rate that are also observed in $_\rho$1GABA$_A$ L9′I (ref. 19) and α7nAChR L9′T[18], support the homology with the 'gain of function' phenotypes found with other pLGICs, and probably cause cell death at pH 7.0–7.2 of the culture medium. The adjacent Glvi Leu8′Ala mutation had no significant effect on Glvi properties (data not shown). These observations suggest that homologous gating mechanisms operate in these distantly related channels. Last, the Glvi channel is not gated by a neurotransmitter but by protons. Within the pLGIC superfamily, some receptors were reported to be modulated by protons[21,22]. In the case of Glvi, several residues close to the binding pocket of eukaryotic pLGICs are titrable (Arg in loop A, B, E and the EDRxE motif in loop C, Fig. 1a, c). These residues might form a cluster whose proton titration would differ in the open and closed conformations of the receptor.

Altogether, the Glvi protein forms cationic channels gated by protons. Unlike eukaryotic pLGICs that undergo fast activation and multi-step desensitization, Glvi shows slow activation and no desensitization. This basic allosteric behaviour may be related to its minimal 'core' structure (Fig. 1d), which probably resembles that of a common ancestor of the superfamily. The additional features of eukaryotic pLGICs may have been subsequently selected to fulfil specific functions. In particular, the cytoplasmic domain contributes to the allosteric transitions[23], to ion permeation[24] possibly at the level of lateral windows[10], and to protein clustering through interaction with rapsyn or gepherin[25]. A simplified architecture has also been reported for the bacterial ionotropic glutamate GluR0, which belongs to a different superfamily of LGICs[26].

Gloeobacter violaceus does not contain thylakoids[2,27]: the inner membrane is the unique site of photosynthesis resulting in primary proton extrusion. Proton gradients are thus expected to be critical to *G. violaceus* metabolism and the Glvi channel may help adaptation to different proton concentrations. This work paves the way for the

study of other prokaryotic pLGICs. In addition, no structure at high resolution of a complete pLGIC has been provided to date. Because of their compact structure and prokaryotic origin, these novel pLGICs seem to be good candidates for structural investigations.

METHODS

HEK cell expression and sedimentation analysis. PCR was used to amplify *glr4197* from the *Gloeobacter violaceus* genome (PCC 7421, Pasteur Culture Collection), and the product was cloned into the PMT_3 vector (HA tag: GYPYDVPDYA). The signal peptide was predicted with Signal P (http://www.cbs.dtu.dk/services/SignalP/). Cell transfection, immunofluorescence and sucrose gradient were performed as previously described[5,28] (see Supplementary Methods).

Electrophysiology. Patch clamp recordings in HEK cells were performed as previously described[15], in cells co-transfected with Glvi and the green fluorescent protein. Normal external solution (called NaCl) contained (in mM): 140 NaCl, 2.8 KCl, 2 $CaCl_2$, 2 $MgCl_2$, 10 Glucose, buffered with 10 HEPES/NaOH pH 7.4 or 10 MES/HCl for lower pHs. In ionic substitution experiments, NaCl was replaced by mannitol to maintain osmolarity (Man), by 140 mM sodium isethionate (Na-Ise), or 140 mM KCl. Patch pipettes contained (in mM) 140 KCl, 5 $MgCl_2$, 5 EGTA and 10 HEPES–NaOH pH 7.3. Expression and two-electrode voltage-clamp in *Xenopus laevis* oocytes were performed as previously described[29]. (See Supplementary Methods.)

***E. coli* expression experiments.** In the MBP–Glvi construct, MBP and Glvi were separated by a short linker (a $(His)_6$ tag was added C-terminal), cloned into the PET 20b vector (Novagen), expressed in the C43 strain (ATCC) at 18 °C, solubilized in (in mM) 20 Tris HCl, 100 NaCl pH 7.4 containing 40 DDM, purified on amylose agarose[30] followed by size exclusion chromatography in 2 DDM buffer to give an elution profile in the expected range for a pentameric oligomer. Electron microscopy (Philips CM12) was performed as previously described[6] (see Supplementary Methods).

Received 7 June; accepted 20 October 2006.
Published online 10 December 2006.

1. Tasneem, A., Iyer, L. M., Jakobsson, E. & Aravind, L. Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. *Genome Biol.* **6,** R4 (2005).
2. Rippka, R., Waterbury, J. & Cohen-Bazire, G. A cyanobacterium which lacks thylakoids. *Arch. Microbiol.* **100,** 419–436 (1974).
3. Gunthorpe, M. J., Smith, G. D., Davis, J. B. & Randall, A. D. Characterisation of a human acid-sensing ion channel (hASIC1a) endogenously expressed in HEK293 cells. *Pflügers Arch.* **442,** 668–674 (2001).
4. Hansen, S. B. *et al.* Tryptophan fluorescence reveals conformational changes in the acetylcholine binding protein. *J. Biol. Chem.* **277,** 41299–41302 (2002).
5. Grutter, T. *et al.* A chimera encoding the fusion of an acetylcholine-binding protein to an ion channel is stabilized in a state close to the desensitized form of ligand-gated ion channels. *C. R. Biol.* **328,** 223–234 (2005).
6. Cartaud, J., Benedetti, E. L., Sobel, A. & Changeux, J. P. A morphological study of the cholinergic receptor protein from *Torpedo marmorata* in its membrane environment and in its detergent-extracted purified form. *J. Cell Sci.* **29,** 313–337 (1978).
7. Taly, A. *et al.* Normal mode analysis suggests a quaternary twist model for the nicotinic receptor gating mechanism. *Biophys. J.* **88,** 3954–3965 (2005).
8. Brejc, K. *et al.* Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* **411,** 269–276 (2001).
9. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423,** 949–955 (2003).
10. Unwin, N. Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *J. Mol. Biol.* **346,** 967–989 (2005).
11. Giraudat, J., Dennis, M., Heidmann, T., Chang, J. Y. & Changeux, J. P. Structure of the high-affinity binding site for noncompetitive blockers of the acetylcholine receptor: serine-262 of the δ subunit is labeled by [3H]chlorpromazine. *Proc. Natl Acad. Sci. USA* **83,** 2719–2723 (1986).
12. Corringer, P. J. *et al.* Mutational analysis of the charge selectivity filter of the α7 nicotinic acetylcholine receptor. *Neuron* **22,** 831–843 (1999).
13. Wilson, G. G. & Karlin, A. The location of the gate in the acetylcholine receptor channel. *Neuron* **20,** 1269–1281 (1998).
14. Schofield, C. M., Trudell, J. R. & Harrison, N. L. Alanine-scanning mutagenesis in the signature disulfide loop of the glycine receptor α1 subunit: critical residues for activation and modulation. *Biochemistry* **43,** 10058–10063 (2004).
15. Grutter, T. *et al.* Molecular tuning of fast gating in pentameric ligand-gated ion channels. *Proc. Natl Acad. Sci. USA* **102,** 18207–18212 (2005).
16. Lee, W. Y. & Sine, S. M. Principal pathway coupling agonist binding to channel gating in nicotinic receptors. *Nature* **438,** 243–247 (2005).
17. England, P. M., Zhang, Y., Dougherty, D. A. & Lester, H. A. Backbone mutations in transmembrane domains of a ligand-gated ion channel: implications for the mechanism of gating. *Cell* **96,** 89–98 (1999).
18. Revah, F. *et al.* Mutations in the channel domain alter desensitization of a neuronal nicotinic receptor. *Nature* **353,** 846–849 (1991).
19. Chang, Y. & Weiss, D. S, Substitutions of the highly conserved M2 leucine create spontaneously opening ρ1 γ-aminobutyric acid receptors. *Mol. Pharmacol.* **53,** 511–523 (1998).
20. Yakel, J. L., Lagrutta, A., Adelman, J. P. & North, R. A. Single amino acid substitution affects desensitization of the 5-hydroxytryptamine type 3 receptor expressed in *Xenopus* oocytes. *Proc. Natl Acad. Sci. USA* **90,** 5030–5033 (1993).
21. Wilkins, M. E., Hosie, A. M. & Smart, T. G. Proton modulation of recombinant GABA(A) receptors: influence of GABA concentration and the β subunit TM2–TM3 domain. *J. Physiol. (Lond.)* **567,** 365–377 (2005).
22. Schnizler, K. *et al.* A novel chloride channel in *Drosophila melanogaster* is inhibited by protons. *J. Biol. Chem.* **280,** 16254–16262 (2005).
23. Hopfield, J. F., Tank, D. W., Greengard, P. & Huganir, R. L. Functional modulation of the nicotinic acetylcholine receptor by tyrosine phosphorylation. *Nature* **336,** 677–680 (1988).
24. Kelley, S. P., Dunlop, J. I., Kirkness, E. F., Lambert, J. J. & Peters, J. A. A cytoplasmic region determines single-channel conductance in 5-HT_3 receptors. *Nature* **424,** 321–324 (2003).
25. Sola, M. *et al.* Structural basis of dynamic glycine receptor clustering by gephyrin. *EMBO J.* **23,** 2510–2519 (2004).
26. Chen, G. Q., Cui, C., Mayer, M. L. & Gouaux, E. Functional characterization of a potassium-selective prokaryotic glutamate receptor. *Nature* **402,** 817–821 (1999).
27. Belkin, S., Mehlhorn, R. J. & Packer, L. Proton gradients in intact cyanobacteria. *Plant Physiol.* **84,** 25–30 (1987).
28. Sallette, J. *et al.* Nicotine upregulates its own receptors through enhanced intracellular maturation. *Neuron* **46,** 595–607 (2005).
29. Paoletti, P., Ascher, P. & Neyton, J. High-affinity zinc inhibition of NMDA NR1–NR2A receptors. *J. Neurosci.* **17,** 5711–5725 (1997).
30. Fischer, M., Corringer, P. J., Schott, K., Bacher, A. & Changeux, J. P. A new method for soluble overexpression of the α7 nAChR extracellular domain. *Proc. Natl Acad. Sci. USA* **98,** 3567–3570 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We wish to thank N. Le Novère, J. L. Popot, P. Delepelaire and C. Beloin for useful assistance, and S. Edelstein for critical reading. This work was supported by the Région Ile de France, the Association Française contre les Myopathies, the Collège de France, the Commission of the European Communities (CEC) and the Association pour la Recherche sur le Cancer.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.-J.C. (pjcorrin@pasteur.fr).

NEUROPHYSIOLOGY

Hodgkin and Huxley model — still standing?

Arising from: B. Naundorf, F. Wolf & M. Volgushev *Nature* **440,** 1060–1063 (2006)

Action potentials in cortical neurons show a variable threshold and a sudden rise in membrane potential at initiation. Naundorf *et al.*[1] fail to explain these features using single- or double-compartment Hodgkin–Huxley-style models, suggesting instead that they could arise from cooperative opening of Na^+ channels, although there is no direct biological evidence to support this. Here we show that these so-called unique features are to be expected from Hodgkin–Huxley models if the spatial geometry and spike initiation properties of cortical neurons are taken into account — it is therefore unnecessary to invoke exotic channel-gating properties as an explanation.

Cortical pyramidal cells initiate spikes in the axon initial segment (AIS) about 30–60 μm from their soma. These spikes then propagate antidromically through the soma and dendrites[2–4]. A well known feature of antidromic spikes is their sudden rise from baseline[5]. These critical properties were not considered by Naundorf *et al.*[1].

We made simultaneous whole-cell recordings from the AIS by patching the cut end of the axon (Fig. 1, legend) and the soma of layer-5 pyramidal neurons *in vitro*[6] during spontaneous spike generation (Fig. 1). Somatic spikes showed a rapid rise, or 'kink', at initiation (Fig. 1a, b) and the slope of the phase plot of spike dV/dt versus V at $dV/dt = 15\ mV\ ms^{-1}$ was $25 \pm 6.8\ ms^{-1}$ (mean ± s.d.; $n = 32$). The phase plots of dV/dt versus V typically revealed a biphasic rise, which was suggestive of two underlying components (Fig. 1b; $n = 30/32$), as observed in many cell types[7,8]. This biphasic component was not evident in the recordings of Naundorf *et al.*[1], although the low peak dV/dt of their recordings indicates that their spikes may not have been fully represented.

Intrasomatic injection of a noisy conductance that mimics the arrival of excitatory and inhibitory synaptic activity[9] resulted in significant variation in the apparent spike threshold ($n = 6$; Fig. 1c, green lines), as observed in the recordings of Naundorf *et al.*[1].

In contrast to somatic spikes, those recorded at the site of spike initiation, the AIS, showed a slower rise ($n = 10$; Fig. 1d, e). The slope of the phase plot of spike dV/dt versus V at $dV/dt = 15\ mV\ ms^{-1}$ was much lower for the AIS ($3.8 \pm 1.7\ ms^{-1}$; $n = 6$; $P<0.01$; Fig. 1d, e) than it was for the soma (Fig. 1a, b). The slow rise at spike initiation in the AIS is not an artefact of our method of axonal recording (Fig. 1, legend). On intrasomatic injection of a noisy conductance that mimics synaptic activity[9], the apparent spike threshold was less variable for the AIS ($n = 6$; Fig. 1f, green lines) than it was for the soma (Fig. 1c).

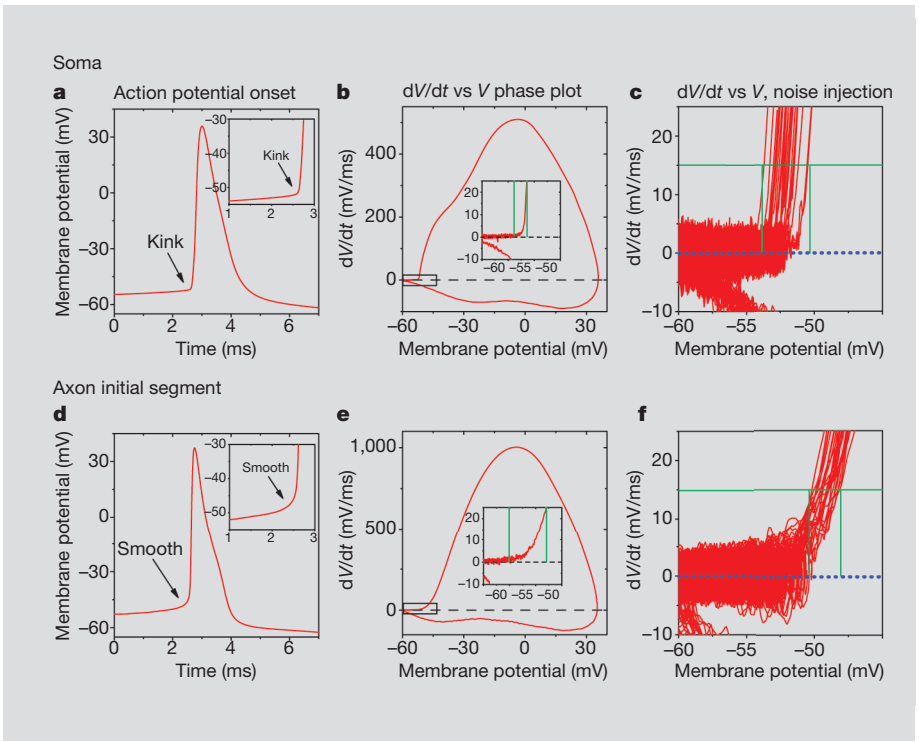


Figure 1 | Properties of spike initiation in the soma and axon of cortical pyramidal cells. a, Somatic spike exhibits a 'kink' at its onset. **b,** Phase plot (dV/dt versus V) and close-up of rapid initiation (inset) of the spike shown in **a**. **c,** Close-up of the phase plot of somatic spike initiation during noisy intrasomatic current injection[9], showing a broad distribution of thresholds (green lines). **d,** Whole-cell axonal recording (50 μm from the soma). **e,** Phase plot of the axonal spike. Note the smoothly rising dV/dt. **f,** Overlay of dV/dt versus V for the onset of axonal spikes, showing lower variability (compare with the soma) of spike threshold (green lines).
Methods. Simultaneous axonal and somatic whole-cell recordings were obtained with the multiclamp 700B amplifier from ferret prefrontal cortical layer-5 pyramidal cells in slices maintained *in vitro* at 36 °C (ref. 6). Spikes shown in **a**, **d**, as well as in **c**, **f**, were recorded simultaneously. Spikes occurred either during spontaneous synaptic activity[6] or in response to the intrasomatic injection of a noisy (10–15 mV) current injection[9]. Whole-cell axonal recordings obtained through patching the cut end of the axon (terminal bleb) do not result in abnormal smoothness of spikes because spikes recorded from distal (> 100 μm) axonal sites also show an onset kink owing to spike propagation (see also www.mccormicklab.org).

Spike initiation in the AIS is mediated by either a high Na^+-channel density in the AIS, as indicated by immunocytochemistry[10,11], or by a lesser density of Na^+ channels, which have a low threshold for activation[12]. Using a previous model of spike initiation in a layer-5 cortical pyramidal cell[13], we adjusted the axonal and somatic densities of Na^+ and K^+ channels until the spike waveform and its derivative were similar to those of our actual recordings (compare Figs 1 and 2).

Our Hodgkin–Huxley model initiated spikes in the AIS that then propagated antidromically through the soma and dendrites, as in real pyramidal cells. At the soma, these spikes showed a rapid rise at initiation (Fig. 2a, b), and the slope of the phase plot for spike initiation at $dV/dt = 15\ mV\ ms^{-1}$ was $21\ ms^{-1}$. Intrasomatic injection of artificial synaptic barrages[9] into the modelled neuron revealed a high variability of apparent spike threshold in the soma (Fig. 2c).

As in the whole-cell recordings, the rise in the model spike at initiation was smoother at the AIS (Fig. 2d, e) than at the soma (Fig. 2a, b). The slope of the phase plot for spike initiation at $dV/dt = 15\ mV\ ms^{-1}$ was considerably lower for the model AIS ($4\ ms^{-1}$) than for the soma, and both were in the range observed in normal cells. Intrasomatic injection of artificial synaptic barrages[9] showed a less variable threshold in

the AIS (Fig. 2f) than in the soma (Fig. 2c), as we found for real neurons (Fig. 1c, f).

We found that several other Hodgkin–Huxley models of cortical pyramidal cells, even one based on a relatively low density of Na^+ conductance in the axon, replicated the 'kink' and variability of somatic spikes (Fig. 2 legend). These features of spike initiation in the soma were dependent on the initiation of spikes in the AIS: increasing the somatic Na^+ conductance to a high level (7.5 nS μm^{-2}) and removing Na^+ conductance from the axon in the model presented here resulted in a loss of the kink at the foot of the spike (soma slope, 4.1 ms^{-1}) and a reduction in spike threshold variability in the soma (results not shown).

Our findings reveal that leading Hodgkin–Huxley models of cortical pyramidal cell spike initiation capture the so-called unique features observed by Naundorf *et al.*[1]. We attribute these features simply to recording from a site that is distant from the site of spike initiation and to the non-uniform distribution of spike properties over the somatic and axonal membrane. The initiation of spikes in the axon that then back-propagate into the soma can result in a rapid change in membrane potential (the kink) at the foot of the somatic spike. The large current supplied by the axonal spike precedes and overlaps with the current supplied by the local generation of the action potential in the soma during the rising phase of the spike. This results in a more rapid rise at the foot of the spike in the soma than would occur if there were no preceding spike in the axon. The apparent high threshold variability with intrasomatic recordings merely results from membrane potential differences between the soma and the actual site of spike initiation, the axon, at the time that spikes are generated. These membrane-potential differences arise from local electrophysiological differences, as well as spatial non-uniformity in synaptic activity. We conclude that the observations made by Naundorf *et al.*[1] are predictable by Hodgkin–Huxley theory and the known physiology of spike initiation[2–4], and that there is no need to invoke exotic interchannel cooperativity to explain their observations.

David A. McCormick*, Yousheng Shu*†, Yuguo Yu*

*Department of Neurobiology, Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, Connecticut 06510, USA
e-mail: david.mccormick@yale.edu

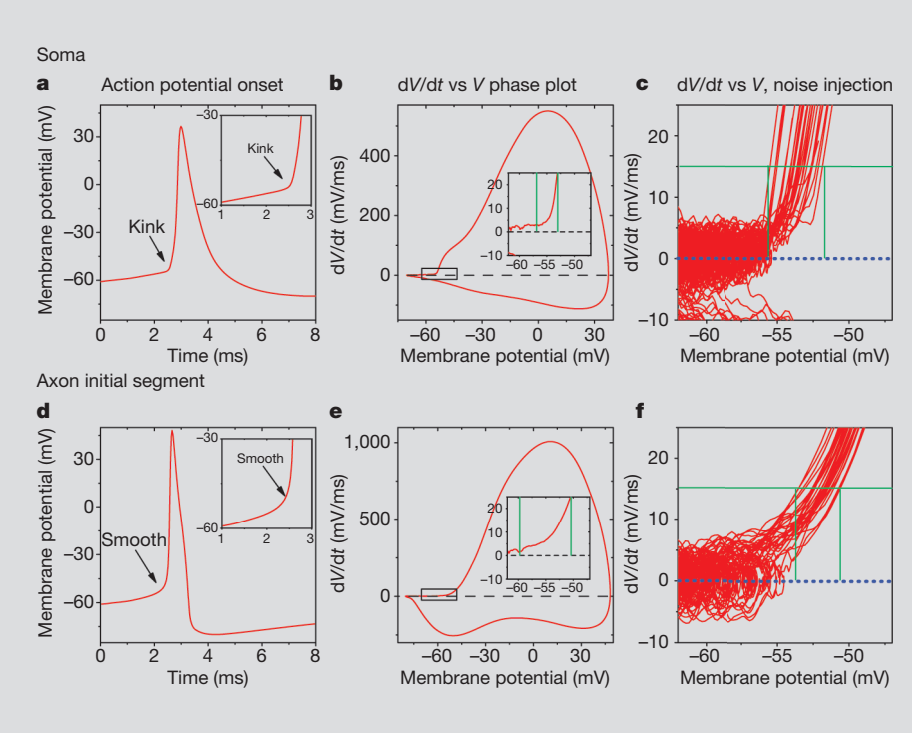


Figure 2 | Hodgkin–Huxley model of a layer-5 cortical pyramidal cell. a, Somatic spike shows a 'kink' at its onset, as in the real neuron. **b,** Phase plot (dV/dt versus V) and close-up of rapid initiation (inset) of the spike shown in **a**. **c,** Close-up of the phase plot of somatic spike during noisy intrasomatic current injection, showing a broad distribution of thresholds (green lines). **d,** Axonal spike (45 μm from the soma). **e,** Phase plot of the axonal spike. Note the smoothly rising dV/dt. **f,** Overlay of dV/dt versus V for the onset of axonal spikes, showing lower variability of spike threshold (green lines).
Methods. Results were obtained from a model layer-5 cortical pyramidal cell[13] with the intrasomatic injection of a 10–15 mV noisy conductance. The model contained the following conductances: soma (Na^+, 0.75 nS μm^{-2}; K^+, 0.15 nS μm^{-2}); axon hillock and initial segment (Na^+, 7.5 nS μm^{-2}; K^+, 1.5 nS μm^{-2}); dendrite (Na^+, 0.1 nS μm^{-2}; K^+, 0.002 nS μm^{-2}; M-current, 0.0003 nS μm^{-2}). Axonal length, 50 μm; soma size, 20 × 30 μm. These parameters were used to match the maximal dV/dt rates, durations and initiation site of spikes in our neurons (Fig. 1). Similar results are obtained from several Hodgkin–Huxley models of cortical pyramidal cells, including those using a high, medium or relatively low density of axonal Na^+ conductance[12–14], and the results from these simulations were well within the range of real cortical cells (see also www.mccormicklab.org).

†Present address: Institute of Neuroscience, Chinese Academy of Science, Shanghai 200031, China

1. Naundorf, B., Wolf, F. & Volgushev, M. *Nature* **440,** 1060–1063 (2006).
2. Palmer, L. M. & Stuart, G. J. *J. Neurosci.* **26,** 1854–1863 (2006).
3. Stuart, G., Schiller, J. & Sakmann, B. *J. Physiol.* **505,** 617–632 (1997).
4. Shu, Y., Duque, A., Yu, Y., Haider, B. & McCormick, D. A. *J. Neurophysiol.* doi:10.1152/jn.00922.2006 (2006).
5. Pare, D., Shink, E., Gaudreau, H., Destexhe, A. & Lang, E. J. *J. Neurophysiol.* **79,** 1450–1460 (1998).
6. Shu, Y., Hasenstaub, A., Duque, A., Yu, Y. & McCormick, D. A. *Nature* **441,** 761–765 (2006).
7. Colbert, C. M. & Johnston, D. *J. Neurosci.* **16,** 6676–6686 (1996).
8. Coombs, J. S., Curtis, D. R. & Eccles, J. C. *J. Physiol.* **139,** 232–249 (1957).
9. Shu, Y., Hasenstaub, A., Badoual, M., Bal, T. & McCormick, D. A. *J. Neurosci.* **23,** 10388–10401 (2003).
10. Inda, M. C., DeFelipe, J. & Munoz, A. *Proc. Natl Acad. Sci. USA* **103,** 2920–2925 (2006).
11. Komada, M. & Soriano, P. *J. Cell Biol.* **156,** 337–348 (2002).
12. Colbert, C. M. & Pan, E. *Nature Neurosci.* **5,** 533–538 (2002).
13. Mainen, Z. F. & Sejnowski, T. J. *Nature* **382,** 363–366 (1996).
14. Baranauskas, G. & Martina, M. *J. Neurosci.* **26,** 671–684 (2006).

doi:10.1038/nature05523

NEUROPHYSIOLOGY

Naundorf et al. *reply*

Replying to: D. A. McCormick, Y. Shu & Y. Yu *Nature* 445, 10.1038/nature05523 (2007)

McCormick *et al.*[1] question whether the rapid onset and highly variable thresholds of action potentials[2] are genuine features of cortical action-potential generators — that is, whether they reflect the voltage-dependence of the underlying sodium currents. Instead, they consider these features to be epiphenomena, reflecting lateral currents from a remote initiation site, and, contrary to direct evidence[3], they assume that sodium currents show canonical kinetics.

Although the lateral current hypothesis of

McCormick *et al.* is superficially plausible, their recordings are inadequate for showing that the dynamics of axonal action-potential initiation conforms to the canonical model. Their so-called axonal recordings are actually obtained from 'blebs' — injury-induced swellings of cut axons on the slice surface. The injured axons, when forming blebs, reorganize their entire cytoskeleton, including the destruction of the sub-membrane spectrin network[4] that integrates sodium channels into the supramolecular machinery of the normal initial segment[5]. As the behaviour of axonal sodium channels is highly sensitive to their cellular environment[6], the smooth action-potential waveforms in the blebs, instead of revealing the true dynamics of action-potential initiation, are more likely to be caused by the disorganized state of the bleb membrane.

The model of McCormick *et al.*[1] does not conform with the known physiology of layer-5 pyramidal cells. Contradicting direct measurements[7,8], it assumes a high ratio of axonal-to-somatic sodium currents. Even with these physiologically unrealistic settings, their model still does not reproduce the experimental data. In their *in vitro* recordings, as in our *in vivo* recordings (Fig. 2 (panels a, c) in ref. 2), somatic action potentials rise almost vertically out of the cloud of subthreshold fluctuations. In their model, however, the range of action-potential onset potentials hardly overlaps with the range of subthreshold fluctuations, being shifted towards more depolarized potentials (Fig. 2 (panel c) in ref. 1). The model of McCormick *et al.* therefore in fact provides further evidence that canonical models are incapable of correctly describing the observed dynamics of action-potential initiation[2,3].

However, McCormick *et al.* highlight an important issue. How can the action-potential dynamics at a remote initiation site be critically probed, when action-potential waveforms recorded from thin processes, such as axons, are likely to be compromised by technical problems[9]? Our analysis identifies an essentially non-invasive approach for addressing this question (see supplementary information of ref. 2). It is based on quantifying the ability of a neuron to phase-lock its spikes to a weak test stimulus in the irregular firing regime[2,10,11].

Theoretical studies indicate that canonical generators of action potentials have a very limited ability to encode high-frequency inputs, showing cut-off frequencies of phase-locking (v_c) that are of the order of their mean firing rate[10,11]. By contrast, models with intrinsically high onset rapidness (r) can show arbitrarily high cut-off frequencies[2,10–12]. If the rapidness of the action-potential onset is genuinely increased by a factor of 10, then cut-off frequencies above 100 Hz are predicted by dimensional analysis ($v_c \propto r$), even for mean firing rates of around 10 Hz. Both *in vivo* and *in vitro* studies have revealed signatures of such fast responses in the neocortex[12,13], supporting genuinely rapid initiation of action potentials in cortical neurons (see also http://www.nld.ds.mpg.de/actionpotentials).

Björn Naundorf*, Fred Wolf*, Maxim Volgushev†
*Max Planck Institute for Dynamics and Self-Organization, Department of Physics and Bernstein Center for Computational Neuroscience, University of Göttingen, 37073 Göttingen, Germany
e-mail: fred@nld.ds.mpg.de
†Department of Neurophysiology, Ruhr-University Bochum, 44780 Bochum, Germany; and Institute of Higher Nervous Activity and Neurophysiology Russian Academy of Science, Moscow 117485, Russia

1. McCormick, D. A., Shu, Y. & Yu, Y. *Nature* **445,** doi:nature05523 (2007).
2. Naundorf, B., Wolf, F. & Volgushev, M. *Nature* **440,** 1060–1063 (2006).
3. Baranauskas, G. & Martina, M. *J. Neurosci.* **26,** 671–684 (2006).
4. Spira, M. E., Oren, R., Dormann, A., Ilouz, N. & Lev, S. *Cell Mol. Neurobiol.* **21,** 591–604 (2002).
5. Lacas-Gervais, S. *et al. J. Cell Biol.* **166,** 983–990 (2004).
6. Rush, A. M., Dib-Hajj, S. D. & Waxman, S. G. *J. Physiol.* **564,** 803–815 (2005).
7. Colbert, C. M. & Pan, E. *Nature Neurosci.* **5,** 533–538 (2002).
8. Ruben, P. C., Ilscher, S. U., Williams, S. R. & Stuart, G. *J. Soc. Neurosci.* abstr. 476.2 (2003).
9. Waters, J., Schaefer, A. & Sakmann, B. *Progr. Biophys. Mol. Biol.* **87,** 145–170 (2005).
10. Fourcaud-Trocme, N., Hansel, D., van Vreeswijk, C. & Brunel, N. *J. Neurosci.* **23,** 11628–11640 (2003).
11. Naundorf, B., Geisel, T. & Wolf, F. *J. Comput. Neurosci.* **18,** 297–309 (2005).
12. Silberberg, G., Bethge, M., Markram, H., Pawelzik, K. & Tsodyks, M. *J. Neurophysiol.* **91,** 704–709 (2004).
13. Williams, P. E., Mechler, F., Gordon, J., Shapley, R. & Hawken, M. J. *J. Neurosci.* **24,** 8278–8288 (2004).

doi: 10.1038/nature05534

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

Are Western universities currently producing too many PhDs or too few? The answer depends largely on whether you are recruiting for a job or looking for one. For recent graduates who are struggling to beat hundreds of other applicants to claim a full-time post, the answer is fairly obvious. But for those in industry who are trawling this sea of talent, the issue is less clear cut. Many recruiters say that they are unable to find the skills they require in the traditional marketplace. Depending on your level of cynicism, this is either an honest assessment or an apologia for outsourcing farther afield.

Data alone seem unable to resolve the issue. A point illustrated last month at a careers discussion hosted by *Naturejobs* at the American Society for Cell Biology's meeting in San Diego, California. Peter Henderson, who reports on scientific workforce issues for the US National Academy of Sciences, noted that there have been at least ten reports covering such issues in the past decade. "Some of the data — even from our own institution — are contradictory," he said.

Cutting through this, Rodney Moses, vice-president for recruitment at reagents firm Invitrogen in Carlsbad, California, explained that he tends to put more faith in his own experiences than in data on the labour market. He told the meeting that he gets a lot of applications from PhD scientists, but that few of them have the requisite business experience. "My ideal candidate would be a PhD with an MBA," Moses said. Failing that, he said, the prospective employee should at least have some understanding of how research skills can translate into the marketplace.

To explore the issues of what recruiters want — and how this matches up with what jobseekers can offer — *Naturejobs* is this week launching a section called Recruiters (see page 124). Every week, this page will explore the job market from the perspective of recruiters in both industry and academia, offering crucial advice and inside information. The first instalment examines the vexed question of supply and demand. We would welcome any contributors, suggestions or comments for this section at naturejobseditor@naturedc.com.

Paul Smaglik, *Naturejobs* editor

Editor: Paul Smaglik
Assistant Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)

UK/Ireland/Italy/RoW:
Loredana Milanese (4944)
Nils Moeller (4953)
Scandinavia/Spain/Portugal:
Evelina Rubio-Morgan (4973)
Germany/Austria/The Netherlands:
Reya Silao (4970)
Online Job Postings:
Matthew Ward (+44 (0) 20 7014 4059)

European Satellite Office
Germany: Patrick Phelan
Tel: +49 89 54 90 57 11
Fax: +49 89 54 90 57 20
e-mail: p.phelan@nature.com

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:**
Catherine Alexander

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
e-mail: a.watanabe@natureasia.com

The inside track from academia and industry

A question of supply and demand

Simply having a PhD may not be enough — you need to marry scientific expertise with the right skills.

Michael Alvarez

During the past few decades, the demand in the labour market for scientists and engineers educated to doctoral level has been a matter of considerable uncertainty and debate. Some say that the number of trainees being produced is too low to keep pace with demand, whereas others believe there are too few opportunities for jobseekers once they have completed their training. So why the discrepancy, and which view is correct? In fact, it is likely that there is some truth in both arguments, and putting them in a broader context helps to reconcile the differences.

PROBLEM: Labour-market projections are complex and, thanks to unforeseeable political and economic changes, must include wide margins of error. For the science and engineering sector, there are significant differences in the supply–demand equation across the various disciplines. This makes it difficult to generalize about future demands for labour.

The picture is further complicated because some employers achieve short-term gains by claiming that there is a shortage of supply. Stating that more PhD trainees are needed, and so encouraging more people to follow that career path, leads to a downward pressure on labour costs and gives employers greater choice when they come to recruit.

WE NEED MORE!
"Unemployment rates for PhDs consistently register below the national average."

In other words, even the most well-intentioned labour-market projections are informed by inherently tenuous data and groups with vested interests.

University laboratories, for example, rely heavily on doctoral and postdoctoral trainees to produce research and share teaching responsibilities, and few departments or principal investigators are eager to volunteer their own labs as a place to make cuts in trainee personnel.

But a strong case can be made for increasing funding and continuing the positive growth in supply, as long as the training experience is tailored to meet actual demands for skills. Put another way, the focus of the supply–demand debate has historically centred on quantity, but to realize the long-term macroeconomic potential for sector growth and to increase employment opportunities for PhD trainees, more qualitative factors must be considered.

SOLUTION: In the past, becoming a good researcher served as adequate preparation for a career as a scientist. Now, although solid research skills are as essential as ever, in many cases they are insufficient when it comes to contributing in a given work setting or advancing one's career. Additional insight into how and where scientists can best use their skills is becoming increasingly necessary, along with opportunities to develop the adjacent skills that will enable a more practical application of a researcher's training.

It is useful to know, for instance, how literature reviews and the formulation of dissertation topics are akin to certain types of market research; to see how the grant-writing process shares similarities with developing project proposals within a business setting; and to recognize the critical roles scientists can have in the exchange of information between researchers and non-scientists in many environments.

Individual trainees may feel, understandably, that they are on their own when it comes to learning new skills and developing professional opportunities in these domains. Most training environments are focused on research, and pay little attention to the practical requirements of the labour market, perhaps owing to the fear that time spent by trainees thinking about something other than their specific discipline might detract from productivity. So, to ensure their own career success, the smartest trainees are beginning to look for evidence that the PhD or postdoctoral experience will offer both excellent research training and career preparation before selecting where to go. And those institutions that adapt to these demands will produce the most capable and competitive scientists of the future.

WE NEED FEWER!
"The annual number of new PhDs that will be needed is much smaller."

Exposure to the various applications of science and an introduction to aspects of other fields is invaluable in the job market and worth pursuing. Trainees with this background can differentiate themselves from other candidates who bring only research skills to the table. It is clear that the job market will benefit if institutions enhance training so that scientists are prepared to contribute in various employment settings, and are assisted in making informed decisions about their career choices and pursuits.

In the bigger picture, this is beneficial to the individual trainees, the organizations that employ them and, in an overall sense, to society for the way it helps ensure people get work that is meaningful, satisfying and necessary. ■

Michael Alvarez is director of Stanford University's School of Medicine Career Center, California.